Learner Persistence by Country

Krishnanand Sagar

2026-01-15

Introduction

Learning Analytics focus on the measurement, collection, analysis and reporting of data about learners and their contexts in order to understand and optimize learning experiences. Massive Open Online Courses (MOOCS) such as those offered by FutureLearn, generate large volumes of learner interaction data that provide valuable opportunities to study engagement and persistence at scale.

Learner engagement is a critical concern for MOOC providers, as low persistent and high dropout rates can limit both educational impact and commercial sustainability. Understanding how engagement patterns vary across learner populations can help providers make informed decisions about course designs, delivery methods, and market focus.

The aim of this report is to Is to analyse learner persistent across countries using FutureLearn course data. The analysis follows CRISP-DM framework and is conducted across two iterative cycles. Cycle 1 focuses on identifying patterns of learner persistence by country, while Cycle 2 builds on these findings to generate stakeholder-oriented recommendations.

Business Understanding (CRISP-DM Cycle 1)

From the perspective of FutureLearn and its partner course providers learner persistence is a key indicator of course effectiveness and learner satisfaction. High dropout rates may signal issues such as accessibility barriers, mismatched learner expectations or insufficient early stage support.

Stakeholders are interested In understanding:

1.Where learners tend to remain engaged throughout a course 2.Where disengagement occurs early 3.How these patterns vary geographically

Research questions - Cycle 1

1.How does letter persistence differ across countries? 2.Which countries show particularly high dropout rates?

The goal of Cycle 1 is descriptive: to identify meaningful differences in persistence that can inform further investigation.

Data Understanding (CRISP-DM Cycle 1)

The analysis uses data from a FutureLearn MOOC, combining learner enrolment information with step-level activity data. Multiple cores runs were included to increase coverage and robustness.

Learners are uniquely identified using a learner_id, Which allows enrollment data to be linked to step level engagement records. The primary focus is on learner progression through course steps.

Key variables used in analysis include:

country: detected learner country learners: number of learners per country avg_max_step: average maximum step reached early_dropout_rate: proportion of learners disengaging within the first two steps

Initial exploration revealed missing country values and highly imbalanced country sample sizes, both of which influenced later preprocessing decisions. After cleaning and aggregation, the dataset contains learner activity records summarized into country level observations. The final analytical data sets consist of approximately 190 country level rows with a small number of numeric variables and categorical variables. The aggregation allows comparison of engagement patterns across countries while reducing noise from individual level variability.

Data Preparation (CRISP-DM Cycle 1)

Data preparation involved cleaning, filtering, and aggregating learner activity records. Learners with missing country information were excluded to ensure valid country level comparisons.

For each learner, the maximum step reached during the course was calculated as a measure of persistence. These learner-level metrics were then grouped by country to produce summary statistics including average persistence and early drop out rates.

All preprocessing was carried out using dplyr within a ProjectTemplate structure. Cleaned datasets were saved and reused across analysis to maintain reproducibility and separation between data preparation and analysis stages. Data quality limitations, particularly small sample sizes from some countries, were acknowledged throughout the analysis.

Exploratory Analysis and Modelling (CRISP-DM Cycle 1)

The first exploratory analysis examined differences in average learner persistence across countries. To avoid misleading comparisons driven by very small samples the analysis focused on the 10 countries with the largest learner populations.

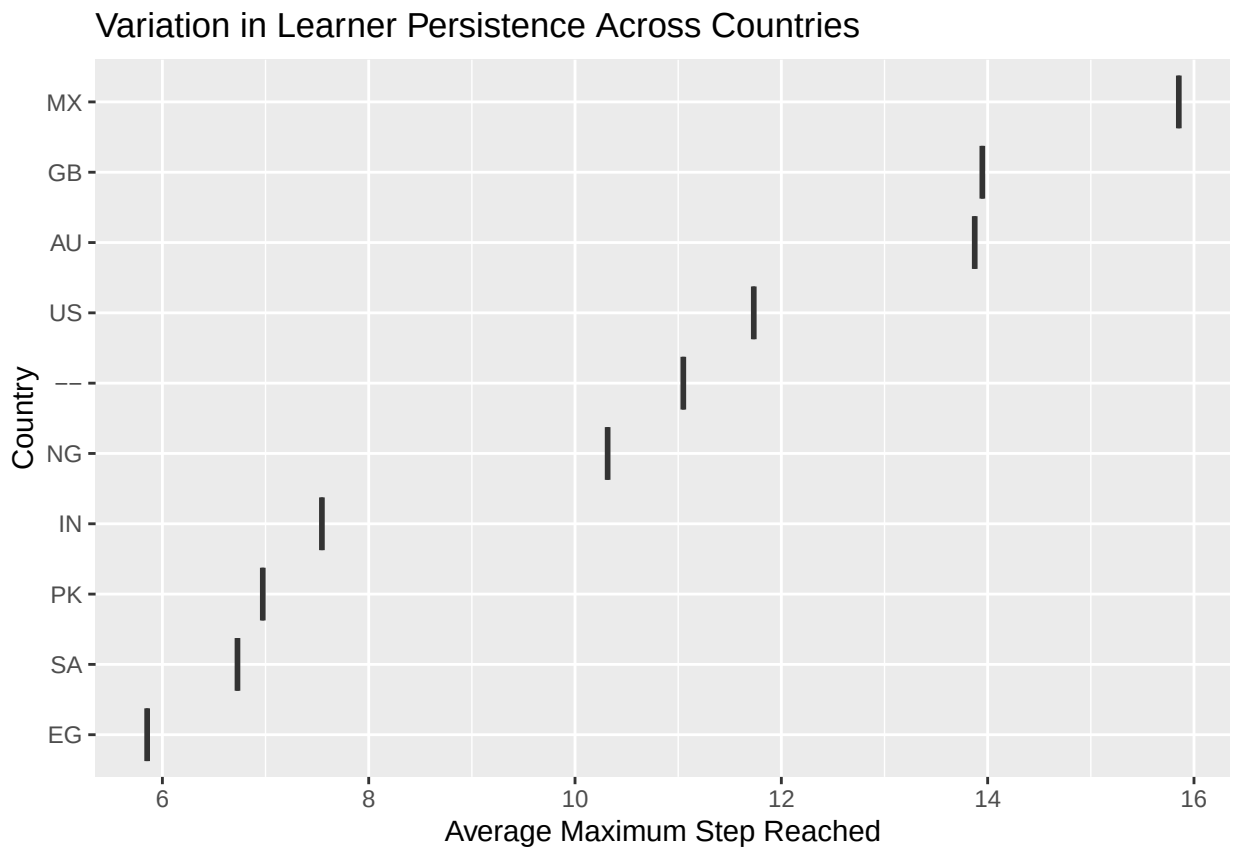
Average maximum step reached was used as descriptive measure of how far learners typically progressed through the course. Visualization using gglot2 highlighted noticeable variation between countries.

A second analysis explored early disengagement by comparing early dropout rates across countries. Learners were classified as early dropouts if they disengaged within the first two steps of the course. Countries were ranked by early dropout rate to highlight where disengagement occurred most rapidly.

These complementary analysis provide insight into both of engagement and early stage attrition.

*Figure 1 shows the average maximum step reached by learners across the ten countries with the most learners. Focusing on populous countries ensures that persistence comparisons are meaningful and not influenced by small sample sizes.

```
country_metrics %>%
slice_max(learners, n = 10) %>%
ggplot(aes(x = reorder(country, avg_max_step), y = avg_max_step)) +
geom_boxplot() +
coord_flip() +
labs(
title = "Variation in Learner Persistence Across Countries",
x = "Country",
y = "Average Maximum Step Reached"
)
```

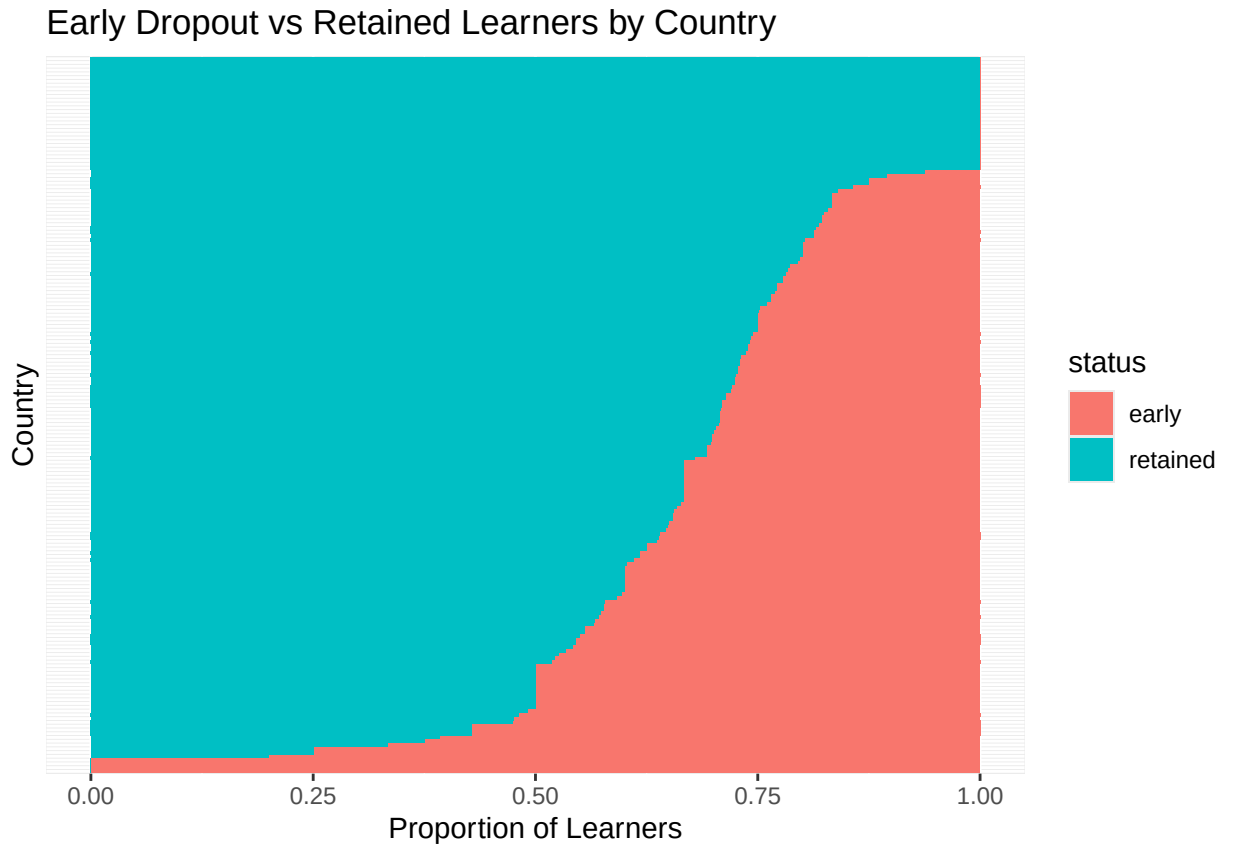


The figure shows clear differences in learner persistence across countries, with some consistently showing higher progression through the course than others.

*Figure 2 breaks down learner engagement into early dropouts and retained learners, allowing a better view of disengagement patterns at the earliest stages of the course.

```
country_metrics %>%
arrange(desc(early_dropout_rate)) %>%
mutate(
country = factor(country, levels = country),
early = early_dropout_rate,
retained = 1 - early_dropout_rate
) %>%
select(country, early, retained) %>%
pivot_longer(
cols = c(early, retained),
names_to = "status",
values_to = "proportion"
) %>%
ggplot(aes(x = country, y = proportion, fill = status)) +
geom_col() +
coord_flip() +
labs(
title = "Early Dropout vs Retained Learners by Country",
x = "Country",
y = "Proportion of Learners"
) +
```

```
theme(
axis.text.y = element_blank(),
axis.ticks.y = element_blank()
)
```



Countries with high early dropout rates show that disengagement happens very early, indicating that initial course structure or accessibility might hinder ongoing participation.

Evaluation and Findings (CRISP-DM Cycle 1)

The results indicate clear variation in learner persistence across countries. Some countries show consistently higher progression through course content, while other experience substantial dropout very early in the course.

From a business perspective:

1.High persistence countries represent stable markets for fully online course delivery. 2.Countries with high early dropout rates may face barriers such as limited access, language challenges or unmet learner expectations.

While cycle one successfully identifies where persistence differs, it does not explain how provider should respond. This motivates a second CRISP-DM cycle focused on actionable stakeholder insights.

These findings meet the success criteria of cycle 1 by clearly identifying differences in learner persistence across countries and highlighting regions with consistently high early dropout rates. The results provide a clear evidence base for refining the business problems and motivate a second CRISP DM cycle focused on actionable stakeholder decisions.

Business Understanding (CRISP-DM Cycle 2)

Cycle 2 reframes the business problem from description to decision support. Building on Cycle 1 findings the goal is to identify countries where low persistence suggest that alternative delivery models may be beneficial.

Research Questions - Cycle 2

1.Which countries can be classified as low, medium or High persistence based on engagement metrics? 2.Which low persistence country should be prioritized for targeted offline or blended learning interventions?

Cycle 2 directly extends Cycle 1 by operationalizing persistence patterns into categories meaningful for stakeholders.

Data Understanding (CRISP-DM Cycle 2)

Cycle 2 uses the same prepared data sets as cycle 1 but introduces derived persistence categories. No new raw data sources were added; instead, the focus shifted to reinterpreting existing metrics in a decision oriented context.

Further inspection highlighted that some countries have extremely small learner counts, reinforcing the need for cautious interpretation of results.

Data Preparation (CRISP-DM Cycle 2)

Additional preparation involved creating rule based persistence categories using quantile thresholds derived from average maximum step and early dropout rates. Countries were classified as low, medium, or high persistence.

These transformations were intentionally simple and transparent to ensure interpretability for nontechnical stakeholders.

Exploratory Analysis and Modelling (CRISP-DM Cycle 2)

Cycle 2 analysis focused on visualizing persistence categories and relationships between engagement depth and early dropout.

Bar chart summarized average learner progression by country and persistence category, Highlighting low persistence regions. Scatter plots further illustrated the relationship between progression and early disengagement across categories.

Rather than predictive modelling, a segmentation based approach was chosen to prioritize clarity and business relevance.

```
country_cycle2
```

```
## # A tibble: 190 x 6
##    country learners avg_max_step median_max_step early_dropout_rate
##    <chr>      <dbl>        <dbl>           <dbl>              <dbl>
##  1 --           382        11.0             8.5              0.280
##  2 AD             1        23              23                0
##  3 AE           113         9.86            7                0.354
##  4 AF            15         7.8             4                0.467
##  5 AG             2        12              12                0
##  6 AL            62         9.31            5                0.290
```

```
##  7 AM            18         5.39             3                0.444
##  8 AO            15        13               11                0
##  9 AR            55         8.62             5                0.345
## 10 AS             1         1                1                1
## # i 180 more rows
## # i 1 more variable: persistence_category <chr>
```

Results and Visualisation - Cycle 2

*Figure 3 summarizes average learner progression by country, grouped by persistence category. This visualisation highlights countries identified as low persistence and aids stakeholders in prioritizing intervention targets.

```
ggplot(
country_cycle2,
aes(x = reorder(country, avg_max_step),
y = avg_max_step,
fill = persistence_category)
) +
geom_col() +
coord_flip() +
labs(
title = "Average Learner Progression by Country",
x = "Country",
y = "Average Maximum Step Reached"
)
```

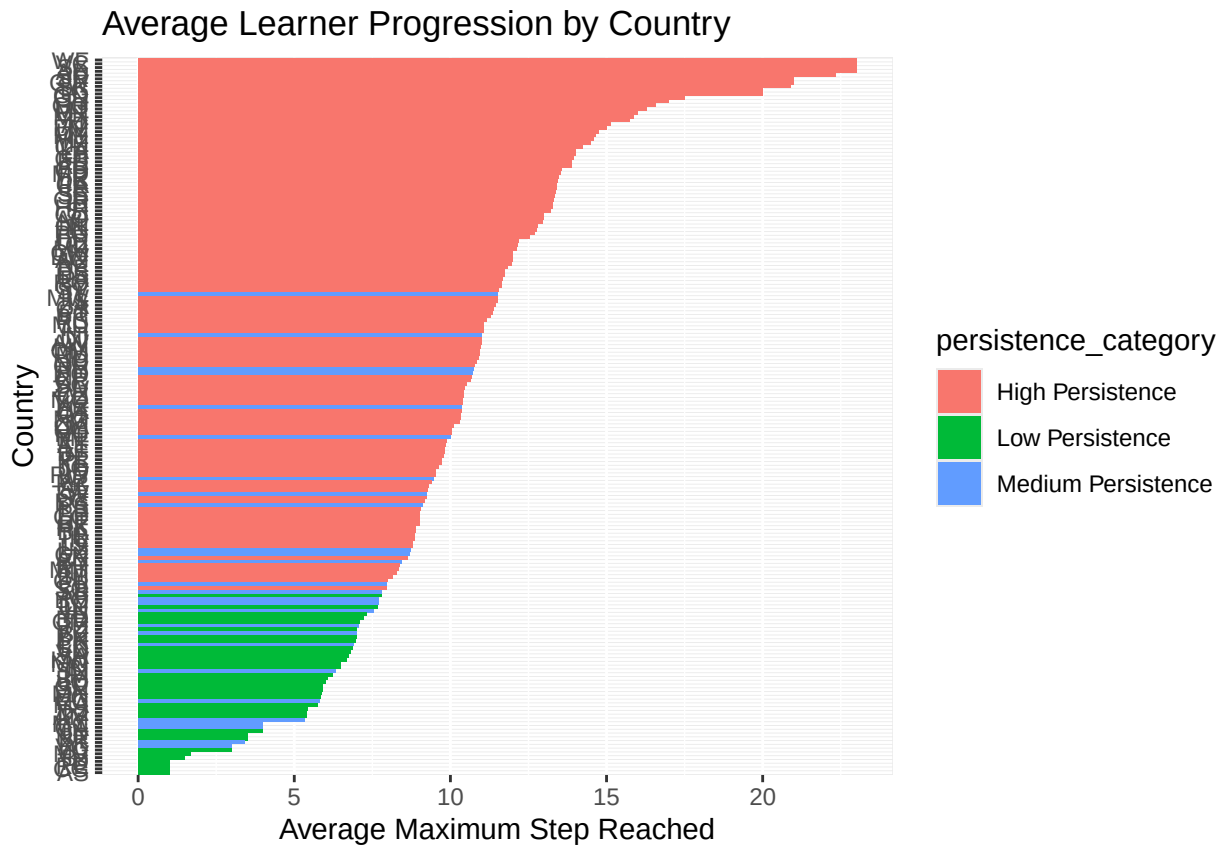


*Figure 4 shows the relationship between average learner progression and early dropout rates, providing another look at engagement dynamics across different persistence categories.

```
ggplot(
country_cycle2,
aes(x = avg_max_step,
y = early_dropout_rate,
color = persistence_category)
) +
geom_point(size = 3) +
labs(
title = "Early Dropout Rate vs Learner Engagement",
x = "Average Maximum Step Reached",
y = "Early Dropout Rate"
)
```

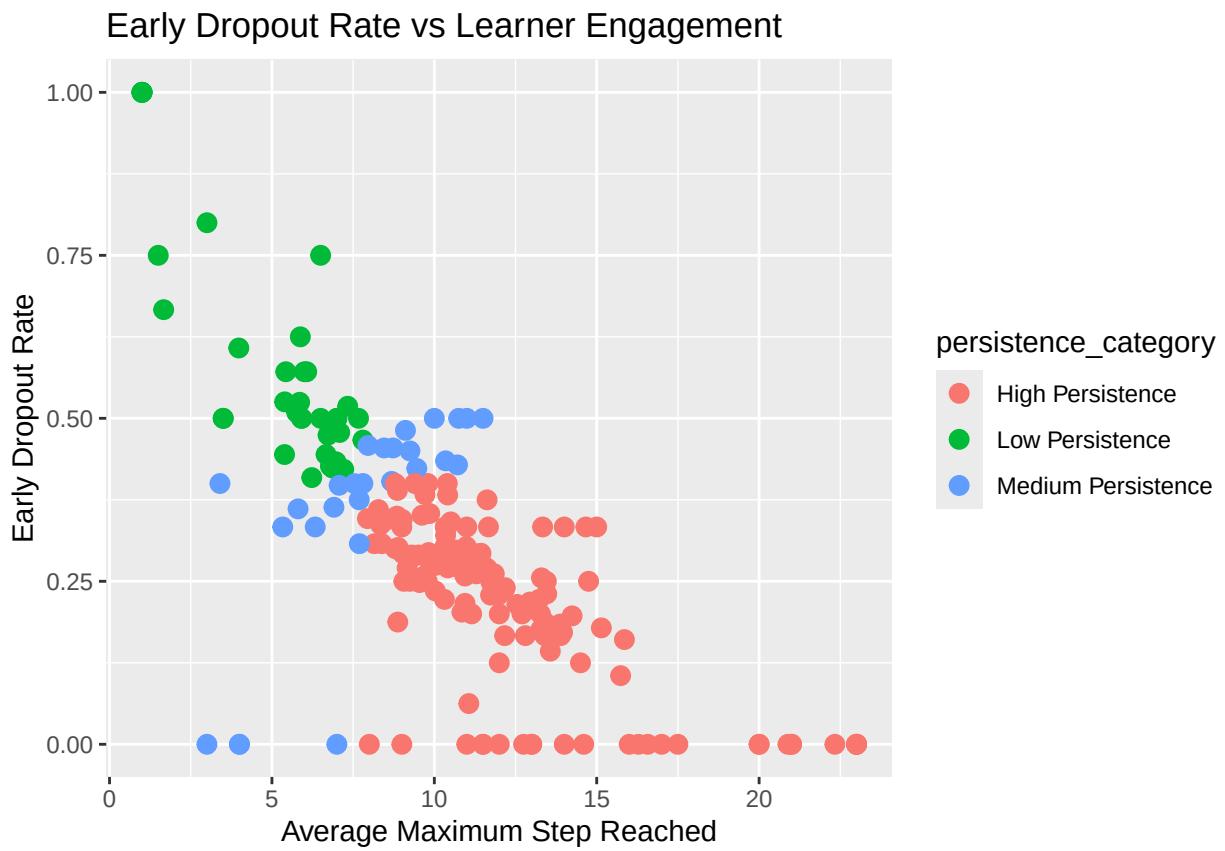


*Table 1 lists countries classified as low persistence, which are priority candidates for targeted offline or blended learning interventions.

```
low_persistence_countries
```

```
## # A tibble: 35 x 6
##    country learners avg_max_step median_max_step early_dropout_rate
##    <chr>      <dbl>        <dbl>           <dbl>               <dbl>
##  1 AS             1         1                  1               1
##  2 CG             1         1                  1               1
##  3 RE             1         1                  1               1
##  4 TD             1         1                  1               1
##  5 LC             5         3                  2               0.8
##  6 BI             4         6.5                1               0.75
##  7 BJ             4         1.5                1               0.75
##  8 MV             3         1.67               1               0.667
##  9 MN             8         5.88               2               0.625
## 10 TN            51         3.98               1               0.608
## # i 25 more rows
## # i 1 more variable: persistence_category <chr>
```

Evaluation and Findings (CRISP-DM Cycle 2)

Countries classified as low persistence exhibit both average progression and high early dropouts. These patterns suggest structural or contextual barriers rather than a lack of learner interest.

Comparing both cycles show a clear progression:

1.Cycle 1 identifies where engagement differs. 2.Cycle 2 translates those differences into prioritized intervention targets.

Despite improvements in stakeholder relevance, limitations related to data granularity and sample size remain.

Deployment and Recommendations

It is suggested that course providers test offline or blended learning initiatives in selected low-persistence countries identified in this analysis. Cycle 1 persistence metrics should be kept as a baseline, and the same analytical method should be reapplied post-intervention to assess changes in learner engagement, dropout rates, and potential revenue impact.

Discussion

Overall, The analysis demonstrates how learner engagement varies substantially across countries within a single MOOC. The two cycle CRISP DM approach allows initial descriptive insights to evolve into practical recommendations.

Methodological choices emphasized transparency, reproducibility and interpretability. Rather than complex models, the analysis prioritized stakeholder friendly metrics aligned into real decision making contexts.

Reproducibility and Best Practice

The analysis follows established data science best practices. ProjectTemplate was used to enforce a clear project structure and separation of concerns. Git version control enable tracking of analytical change over time, while renv ensured consistent package versions.

The report was produced using R Markdown, allowing code, results, and interpretation to remain tightly integrated. These practices reflect real world analytical workflows and support reproducibility and auditability.

Conclusion

This report analyzed learner persistence by country using FutureLearn MOOC data across two CRISP DM cycles. The findings highlight substantial variation in engagement patterns and identify countries where early disengagement is particularly pronounced.

By progressing from descriptive analysis to actionable segmentation, the report demonstrates how learning analytics can support strategic decisions for course providers.

Limitations and Future Work

The analysis is limited by missing country information, imbalanced sample sizes and use of step completion as proxy for engagement. Future work could incorporate demographic variables, temporal engagement patterns or post intervention data to assess the effectiveness of blended learning strategies.