Clustering and Regression Analysis on Student Performance Data from the MathE Platform

Abstract

This report investigates student performance using data collected from the MathE platform, which includes interactions with math-related questions across countries. Our study focuses on understanding behavioral patterns using clustering and predicting student correctness using linear regression. Data preprocessing, feature engineering, and detailed exploratory analysis were employed to build a foundation for the modeling techniques.

1. Introduction

Machine learning in educational contexts offers the ability to model and predict student performance, inform curriculum design, and personalize learning. The MathE dataset provides structured data from thousands of student-question interactions, making it an ideal case for unsupervised and supervised learning applications.

In this study, we apply K-Means clustering to group similar student behaviors and Linear regression to model the likelihood of answering correctly. The goal is to interpret learning patterns and assess the predictive capacity of basic models for student outcomes.

2. Data Description and Initial Exploration

The dataset comprises 9,546 records and 8 features, including Student ID, Student Country, Question ID, Type of Answer, Question Level, Topic, Subtopic, and Keywords.

Type of Answer is nearly balanced (mean ≈ 0.47), making it a fair prediction target. No missing values or structural anomalies were detected. The data showed broad coverage across topics (14) and countries (8).

3. Feature Engineering

Record Count

This feature was derived by grouping the dataset across all available columns, including Student ID, Question ID, Topic, Subtopic, Keywords, and most importantly, Type of Answer. The resulting Record Count represents the number of times an identical student-question-topic-answer combination appears in the dataset.

Purpose and Interpretation

The motivation behind this feature was to capture repeated student behavior on the same question with the same outcome:

- If Type of Answer = 1 and Record Count = 3, it means the student answered that question correctly three times.
- If Type of Answer = 0 and Record Count = 2, it means the student answered the same question incorrectly twice.

Why It Matters

- This feature provides a quantitative measure of learning consistency.
- It distinguishes one-time success/failure from persistent behavior.
- It embeds behavioral and performance frequency into the dataset.

Impact on Modeling

Including Record Count allows clustering algorithms to better separate students based on stability of performance and gives regression models a stronger understanding of repeated behavior patterns.

4. Data Transformation

4.1 Label Encoding

The dataset includes categorical variables such as Student Country, Topic, and Subtopic. Since machine learning algorithms generally require numerical inputs, these features were transformed using Label Encoding:

Each unique category was mapped to an integer.
This encoding method preserves no inherent ordering but ensures the data is model-ready.

4.2 Feature Scaling (Commented Out)

Although normalization using StandardScaler was initially considered for the Record Count feature, the final version of the code does not apply scaling, as this step was commented out.

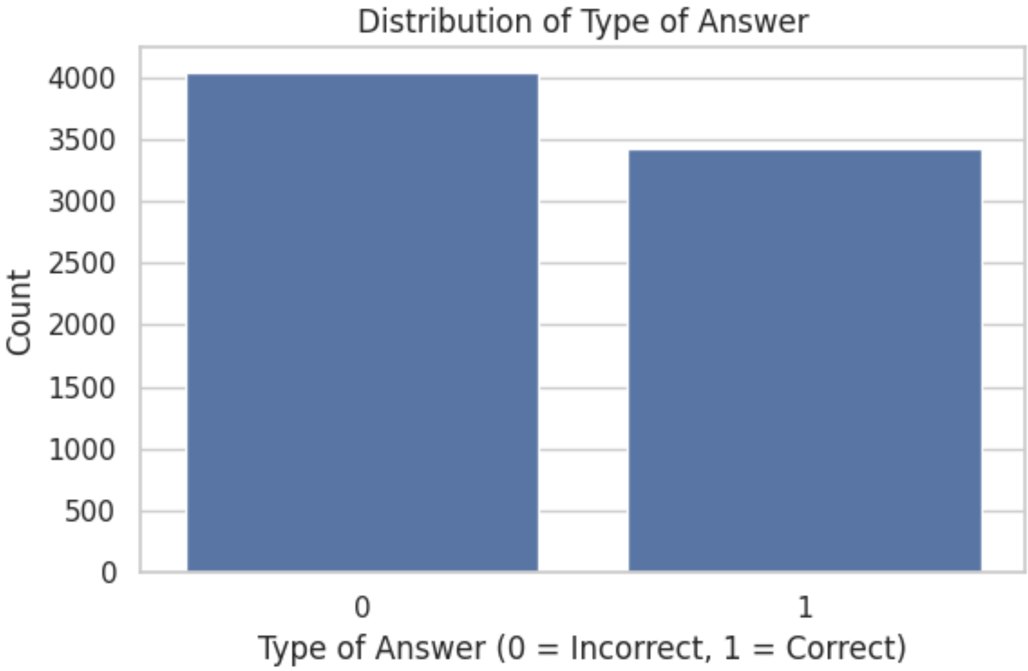
This suggests either scaling was not necessary for your modeling strategy or was omitted based on preliminary experimentation.

5. Visualizing and Addressing Imbalances in the MathE Dataset

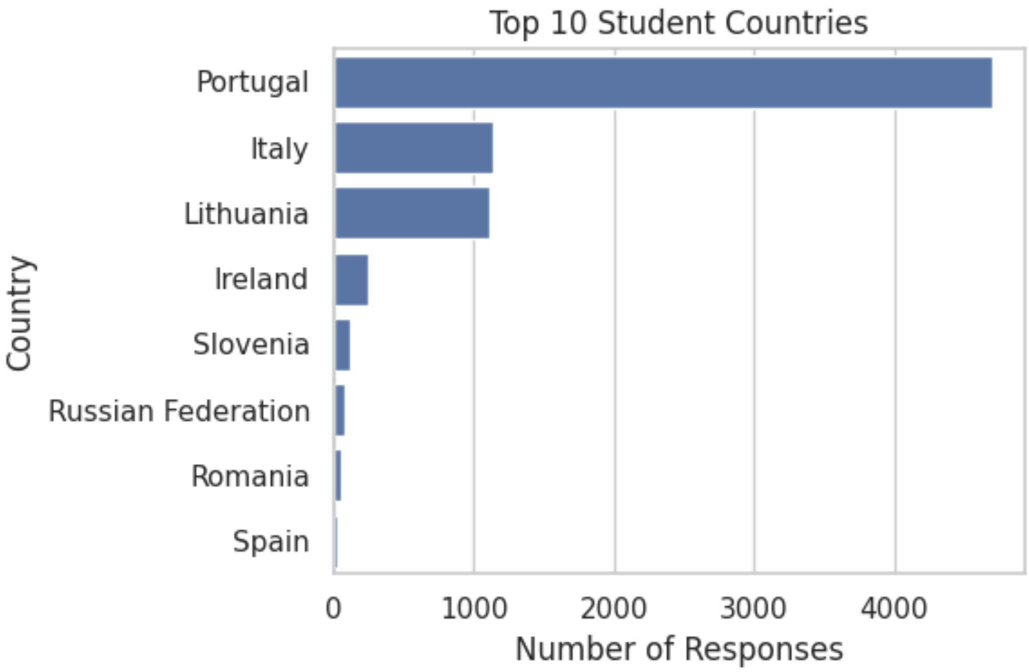
During the exploratory data analysis phase, several visualizations highlighted imbalances in the dataset, such as unequal distributions of answers across countries, topics, and question levels. Addressing these issues is essential for developing fair and generalizable models.

1. Visual Evidence of Imbalance

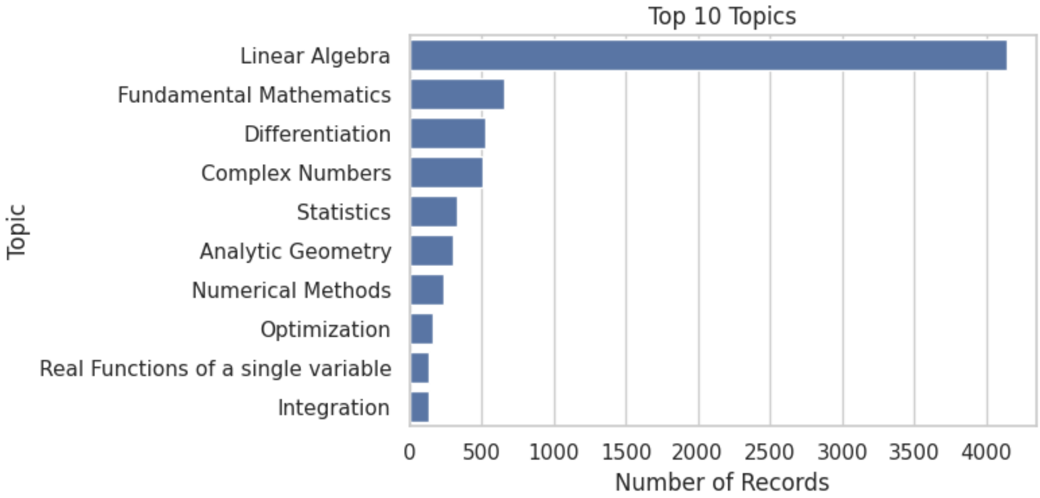
- Imbalance in Correct vs. Incorrect Answers



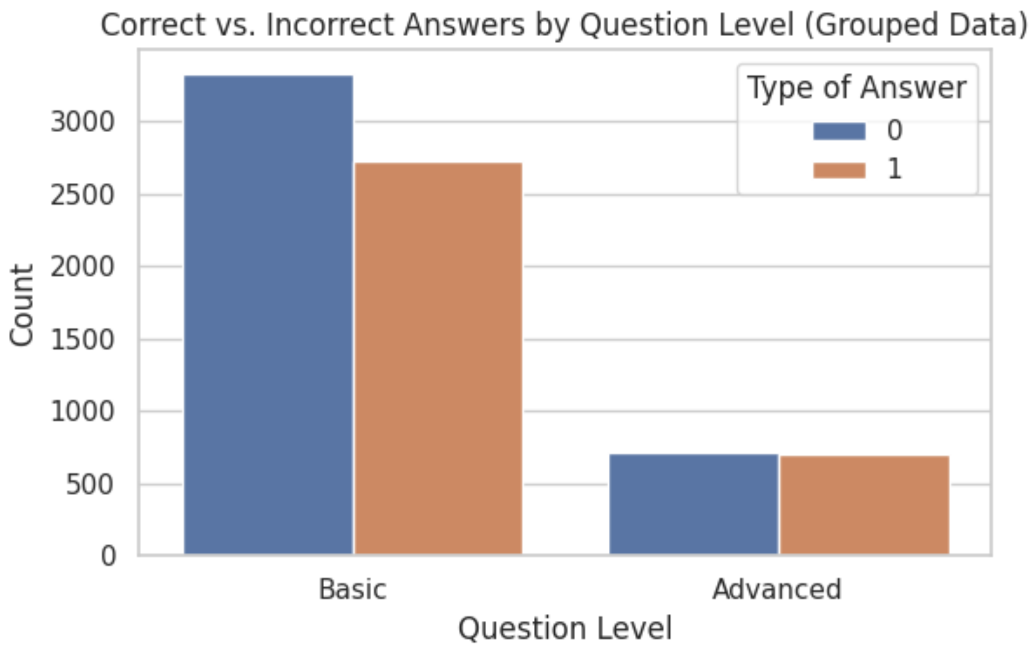
- Overrepresentation of Countries



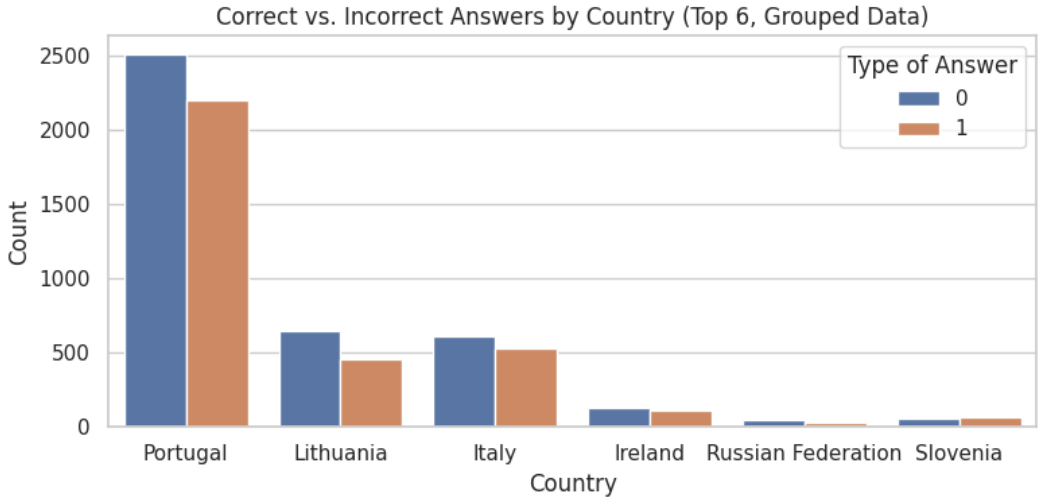
- Overrepresentation of Topics



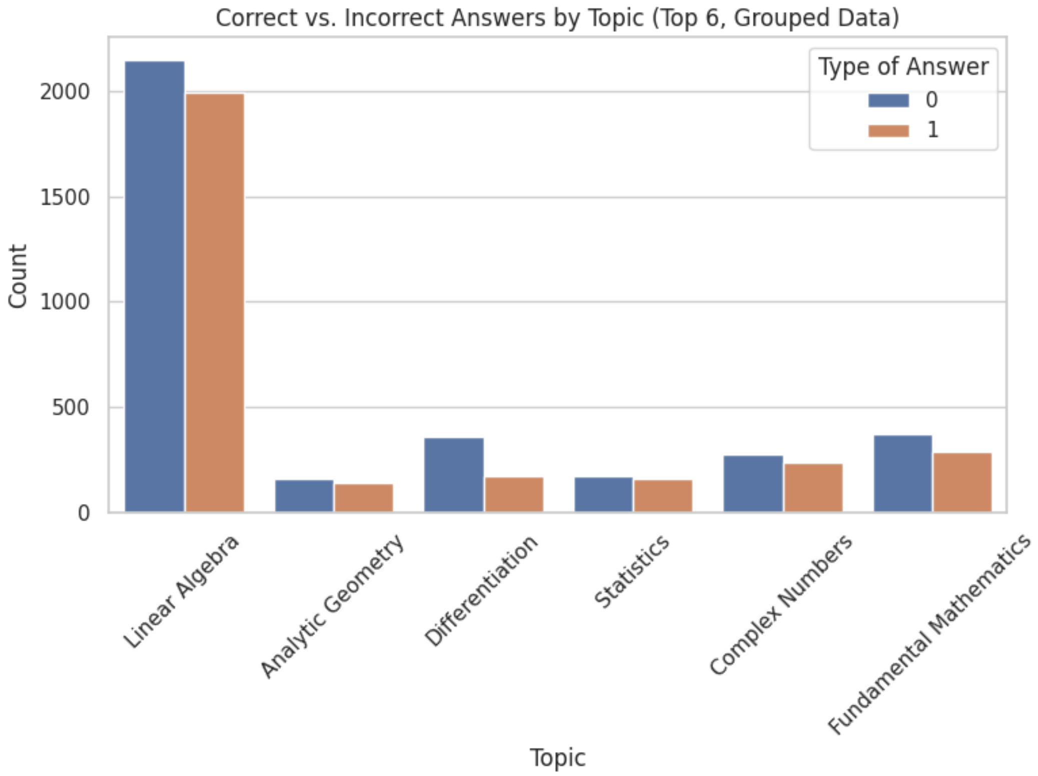
- Question Level Answer Distribution



- Country-wise Answer Imbalance



- Topic-wise Answer Imbalance



2. Addressing the Imbalance

To correct these imbalances, stratified sampling was applied using a composite key formed from the `Student Country`, `Topic`, and `Type of Answer`. This ensured proportional representation across combinations of categories in both the training and test sets.

Code Used:

```
df_model['Stratify_Key'] = (
   df_model['Student Country'].astype(str) + "_" +
   df_model['Topic'].astype(str) + "_" +
   df_model['Type of Answer'].astype(str)
)
```

Only those combinations with more than one record were used to ensure stable stratification. This process ensured that the distribution of correct and incorrect responses remained balanced within each category group across both training and testing datasets.

6. Clustering Analysis (K-Means)

A K-Means clustering algorithm was applied to the dataset with k = 4. This value was selected based on the silhouette score, which was calculated as 0.58. This indicates moderately strong cluster separation, suggesting that the student interaction data naturally forms four distinguishable behavioral groups.

Clusters were created based on encoded features such as Topic, Country, Question Level, and Record Count. The analysis revealed:
- Cluster 0: High correct response rate.
- Cluster 1: More incorrect responses.
- Cluster 2: Average mixed performers.
- Cluster 3: Outlier patterns or specific behavior groups.

The clustering provided meaningful insights into patterns of learning across countries and topics.

7. Linear Regression Evaluation

Linear regression was applied to predict whether a student would answer a question correctly (`Type of Answer`). However, since the target is binary (0 or 1), linear regression is not naturally suited for this task.

Performance Metrics:

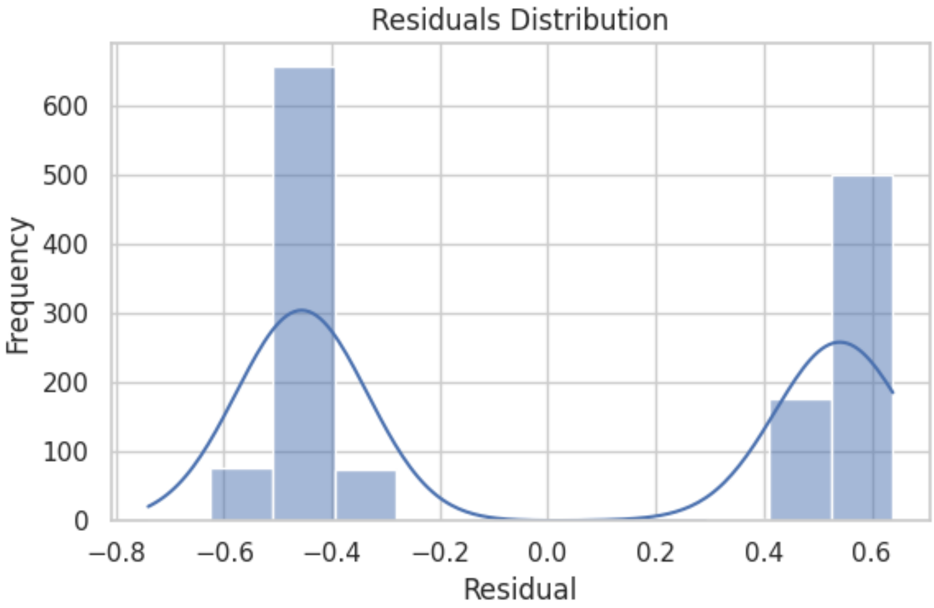
• RMSE: 0.4976

• R^2 Score: 0.0029

The RMSE is close to 0.5, indicating prediction is near random chance, and the R^2 score is nearly zero, showing the model explains less than 1% of the target's variance. This confirms poor model fit.

Residuals Analysis

The residuals distribution below shows a clear bimodal shape, suggesting the model often either overpredicts or underpredicts. This further confirms that linear regression is not appropriate for binary classification.



Conclusion and Recommendation

Given the binary nature of the prediction task and poor regression performance, linear regression is not an appropriate model. It is recommended to use a classification model such as:

- Logistic Regression
- Random Forest Classifier
- Gradient Boosting

These are better suited to handle binary outputs and will provide more accurate and interpretable predictions.

8. Conclusions and Insights

The clustering analysis revealed clear groupings within the data, validating the presence of distinct behavioral patterns among students. A silhouette score of 0.58 for k = 4 supports the quality of cluster separation.

On the regression side, linear regression proved unsuitable due to the binary nature of the target variable. While it helped establish a baseline, the residual plot and R^2 score near zero confirmed its ineffectiveness.

Even after switching to classification algorithms, including Logistic Regression, Random

Forest, and Gradient Boosting, the models underperformed:
- R^2 scores were negative, indicating they performed worse than a constant predictor.
- RMSE values ranged from 0.58 to 0.66, suggesting limited predictive capability.

These results emphasize that while some clustering patterns exist, the feature set used for prediction may lack strong correlation with the target variable.

9. Future Work

1. **Use Classification Metrics:** Since the target is binary, future evaluations should use metrics like Accuracy, Precision, Recall, F1-Score, and AUC.
2. **Advanced Feature Engineering:** Introduce new behavioral features (e.g., student consistency, question attempt history, time on question).
3. **Balance the Dataset:** Explore class balancing techniques such as SMOTE or undersampling to handle slight imbalance.
4. **Model Tuning:** Perform hyperparameter tuning on models like Random Forest or Gradient Boosting to potentially improve performance.
5. **Try Deep Learning Models:** If data volume supports it, use neural networks or embeddings to capture non-linear patterns.

These steps may enhance model performance and provide deeper insight into student learning patterns.

10. References

1. Scikit-learn Documentation – https://scikit-learn.org

2. MathE Dataset (2024 release)

3. Kotsiantis et al., "Data Preprocessing for Supervised Learning", IJCS Vol. 1, No. 1, 2006

To provide additional behavioral context to the dataset, a new column called Record Count was introduced.