Dataset for Assessing Mathematics Learning in Higher Education

Edoardo Clerici [2143888], Omar Dernjani [2146259], Emanuele Foschini [2156935]

22 gennaio 2026

Indice

1 Introduction

We have chosen to analyze the dataset made by *MathE*. This dataset contains student answers to mathematics questions on various topics, like linear algebra, fundamental mathematics and others. We have aimed to analyze the student answers, sorted by the question topic, question difficult level, and students nation; all this to finally achieve different predictive models to classify the correctness of the answers . As long as we are talking of a classification problem, this means that it is a binary problem, where the binary outputs 0 and 1 are represented by the incorrectness or the correctness of the answer. The creators of this dataset have collected data by using two different methods:

- **Data collection from surveys:** in 2020, anonymous surveys were given to lecturers and students registered on the MathE platform. In total, there were 56 students and teachers that answered the anonymous surveys. They have to evaluate the impact of MathE platform.
- **Data collection from students results:** the data was extracted from the MathE platform's database. Data collection was given by the results of 122 students. The results in particular are the answered that students have done. Students have answered to 2973 questions to 24 topics, including subtopics, in the year 2020. Data collected included the right and wrong answers sorted by difficulty level and by topic.

2 Data description

In this dataset we have 8 different features and 9546 instances.

- **Student ID:** each student has an ID. The dataset has 372 students.
- **ID question:** each question has an ID. The dataset has 833 questions.
- **Student country:** contains the country where the student is studying. There are 8 countries, which are Portugal, Lithuania, Italy, Ireland, Romania, Russia, Spain, and Slovenia.
- **Type of answer:** contains the type of answer of each student. Correct answers correspond to 1; incorrect answers correspond to 0.
- **Question level:** questions are divided in two levels: basic and advanced.
- **Topic:** describe the topic of the question. There are various topics, like Linear Algebra, Fundamental Mathematics, Graph Theory, Differentiation, Integration, Analytic Geometry, Complex Numbers, Differential Equations, Statistic, Real Functions of a Single, Variable, Probability, Optimization, Set Theory, Numerical Methods.
- **Subtopic:** describe the subtopic of the question. There are 24 subtopics.
- **Question keywords:** contains the keywords that belong to the question.

Tabella 1: Description of the dataset variables

Variable	Type	Description
Student Country	Categorical	Student's country
Question Level	Categorical	Difficulty level of the question
Topic	Categorical	Topic of the question
Subtopic	Categorical	Subtopic of the question
Keywords	Categorical	Keywords associated with the question
Question ID	Numerical	Unique identifier of the question
Student ID	Numerical	Unique identifier of the student
Type of Answer	Target (binary)	Answer of the question (0 = incorrect, 1 = correct)

3 Data pre-processing

In this phase of the project, we have chosen to adopt two different strategies to processing data. We have used MLB, which is the acronym of MultiLabel Binarizer, and OneHotEncoding. The **target** variable is represented by "Type of Answer".

- **MultiLabel Binarizer and OneHotEncoding:** categorical variables were converted by using OneHotEncoding, while the feature **Keywords** is divided in string lists and then transformed by using MultiLabel Binarizer.
- **OneHotEncoding:** in this case, all the categorical variables are transformed by using OneHotEncoding.

We divided the dataset in two sets:

- **Training set:** which is the 80% of the dataset.
- **Test set:** which is the 20% of the dataset.

4 Models

We evaluated data by utilizing different types of models, like linear models, generative probabilistic models, tree-type models, and discriminant models, to have a full comprehension of how data is working on different models. Each model has been instantiated for MLB and OHE.

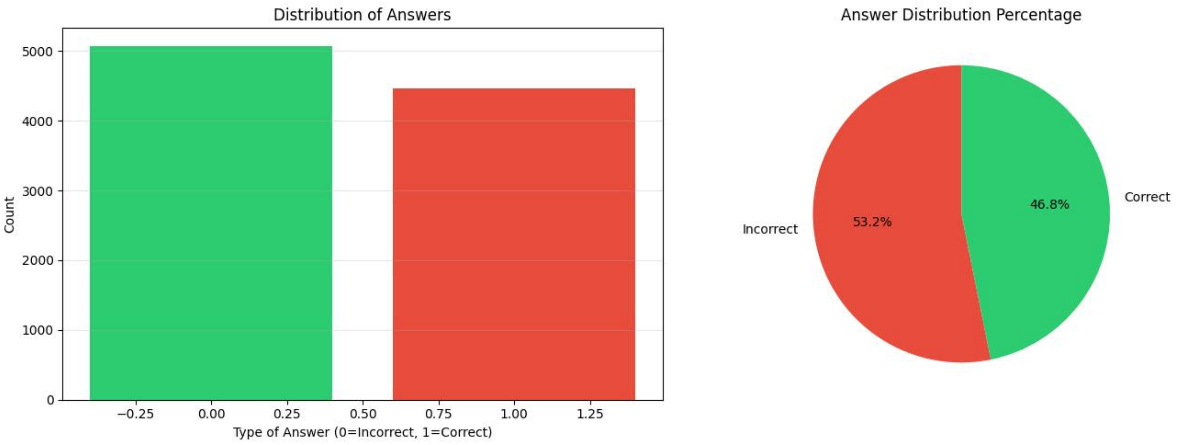
- **Logistic Regression:** this model is used for classification problem (like this one), and is used to predict a categorical outcome, by using a probabilistic method with the sigmoid function. The outcome estimated by the model is a value between 0 and 1.
- **Multinomial Naive Bayes and Bernoulli Naive Bayes:** Multinomial Naive Bayes is useful in text classification problems with discrete features. Is useful to construct the frequencies of words. Bernoulli Naive Bayes is a variant of Multinomial Naive Bayes. It is used in text classification; in this case, the model evaluate the presence or the absence of a word.
- **K-Nearest Neighbors:** this model classify a new data by find a class of its k-nearest in the training dataset.
- **Decision Tree:** it is useful in case of non linear relationships, where it has the function of predict the value of a target variable, by using decision rules.

- **Linear discriminant analysis and Quadratic discriminant analysis:** Linear discriminant analysis does the data distribution for classes by using Bayes' theorem, by finding the optimal linear combination of features in order to separate classes. Although Quadratic linear analysis is similar to LDA, it has some differences. Firstly, in Quadratic linear analysis is possible that the classes' covariance is not identical. Then, in Quadratic linear analysis, decision boundary are quadratic; also this model is more flexible than LDA, but the risk of overfitting is higher.

5 EDA: Exploratory Data Analysis

During the construction of EDA, we managed to show more graphics as possible, to offer the reader a full knowledge of the data, and the different analysis we made. Each graphic is followed by a description of what it represents, so let's start to see in order the plotted graphics.

5.1 Distribution answers and answer distribution percentage



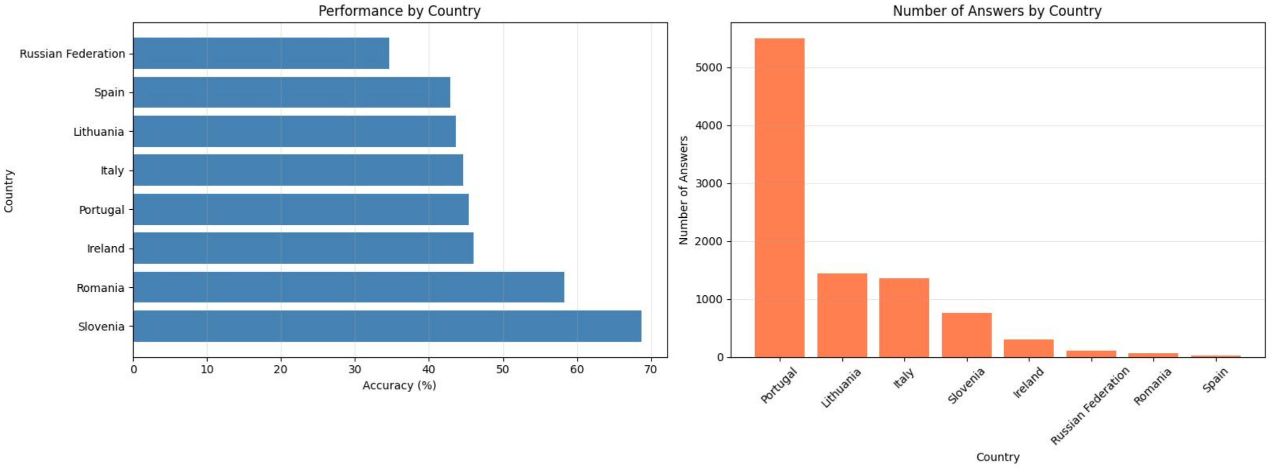
As we can see from the first graphic, the distribution of the wrong answers, which is the red one, is higher than the distribution of the right answers. Especially, in the answer distribution percentage graphic, it's displayed that the 53.2% of the answers are uncorrected, while the remainder percentage, 46.8% is represented by the right answers.

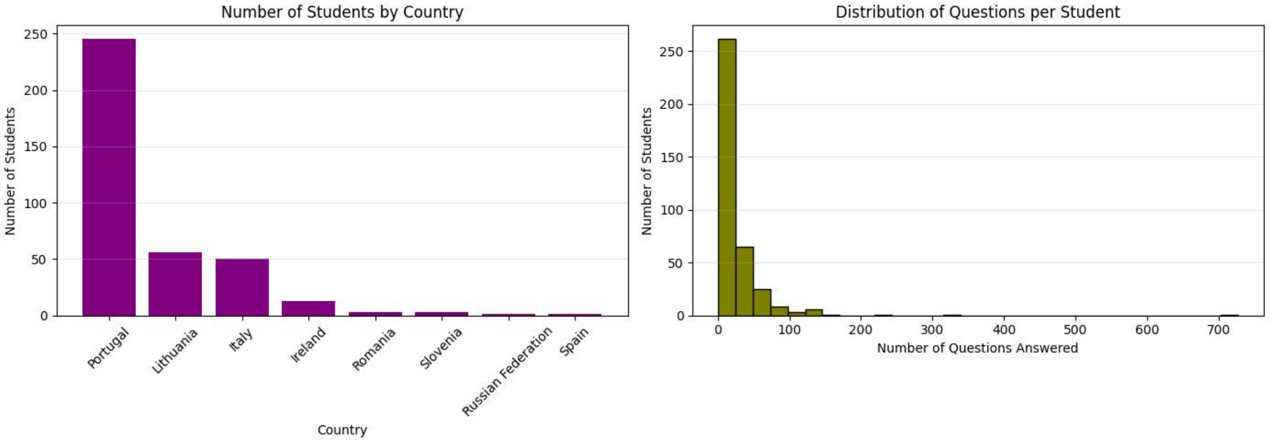
5.2 Performance and number of answers by country

These graphics show the performance by country, by counting for each country the accuracy percentage and the numbers of answers by country. In the first graphic we see that countries like Slovenia and Romania have higher accuracy, even though they are poor of number of answers. Portugal is one of the best countries, because it's the fourth higher country for accuracy percentage, and it's the first country for the number of answers given from students, while countries like Spain or Russian Federation are the worst in the grid.

5.3 Number of students by country and distribution of questions per student

These graphics show the amount of students that were involved by country; as we can see, Portugal leads the graphic with almost 250 students involved, Spain is the last country for

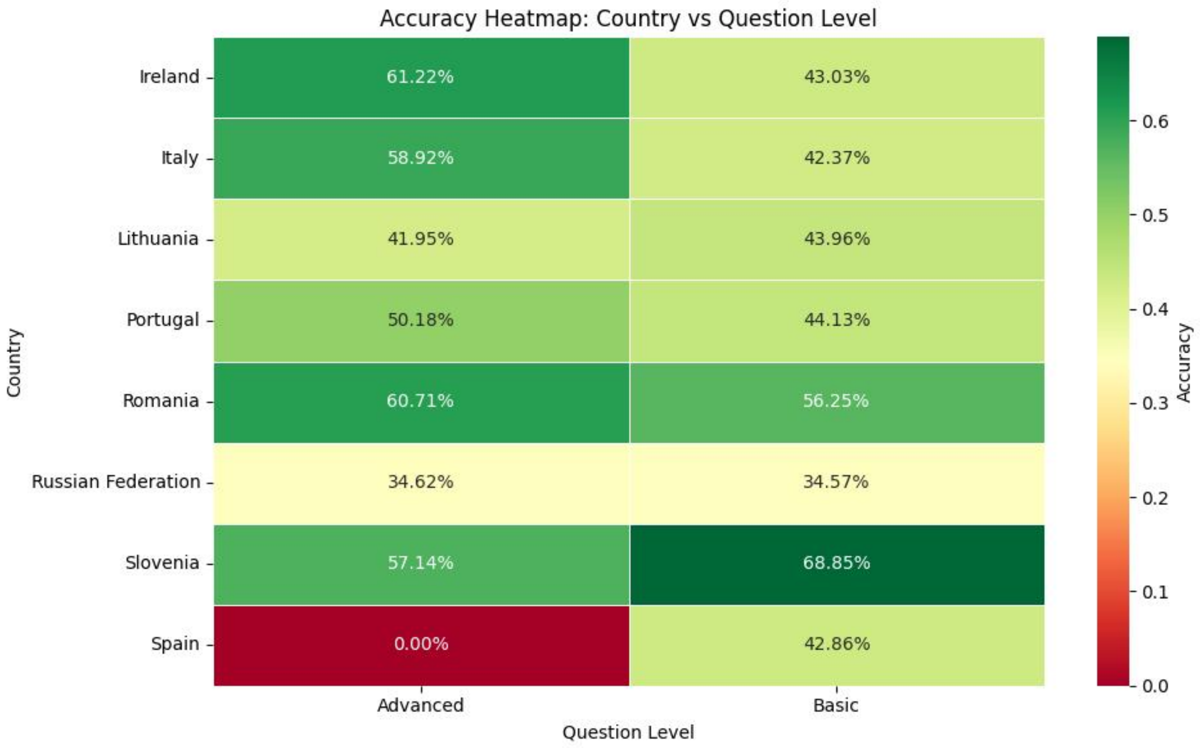




number of students. The second frequencies histogram tells us that the majority of students have answered not many questions; few students have answered more than 100 questions.

5.4 Accuracy heatmap: Country vs question level

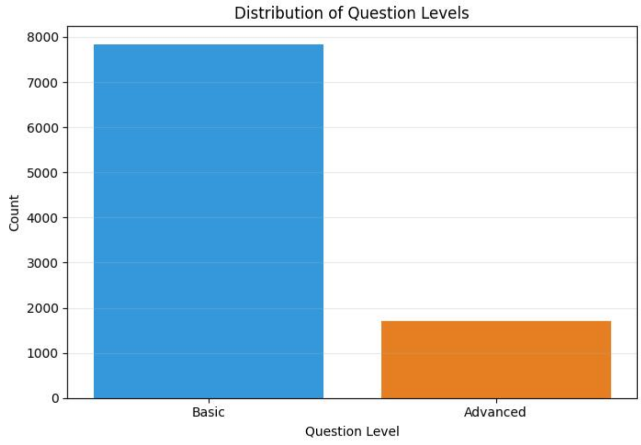
This graphic shows the percentages obtained by each country in advanced or basic questions using a heatmap.

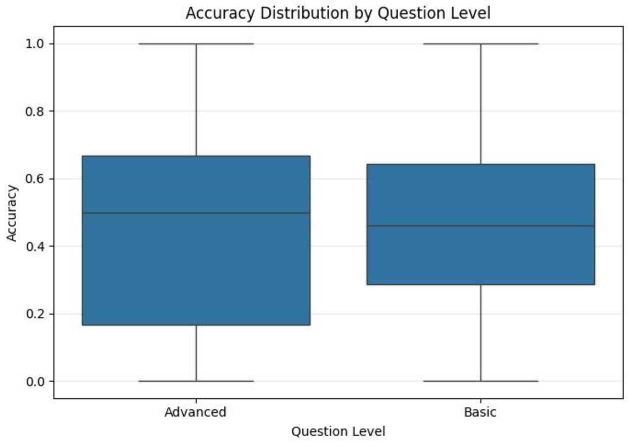


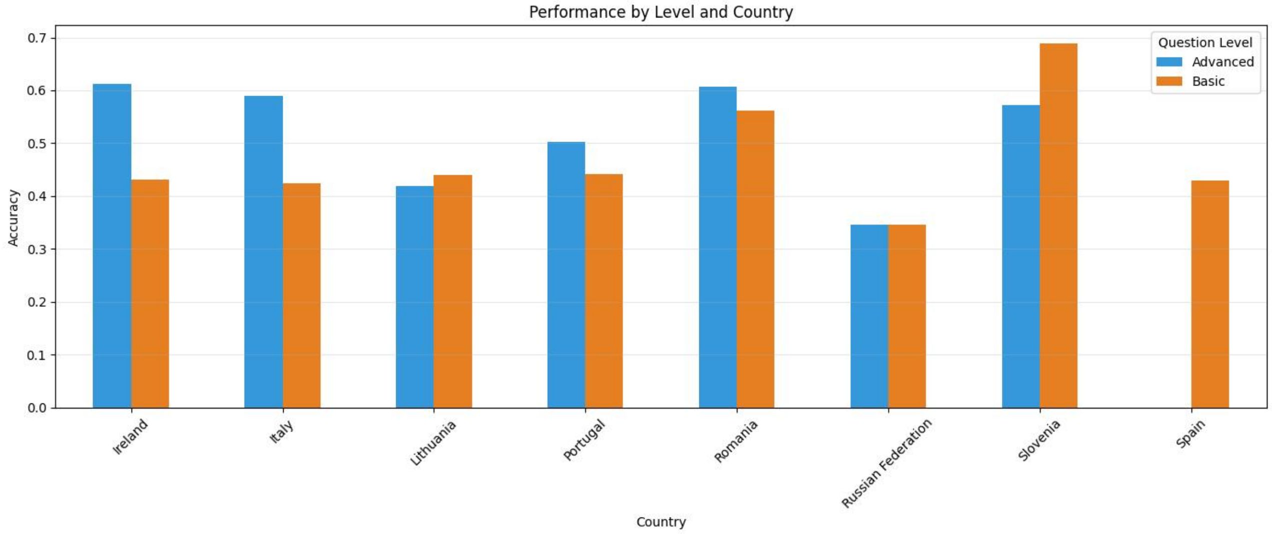
- **Basic question**: the best country for basic question is Slovenia with 68.85%, the worst country is Russian Federation with 34.57%
- **Advanced question**: the best country for advanced question is Ireland with 31.22%, the worst country is Spain with 0.00%

Overall, as we said before, countries like Spain and Russian Federation have worst percentage, Romania and Slovenia have best percentage, and the rest of the countries are in the middle of these two groups. We have created another graphic that shows the performance by level for each country that is correlated to the previous one.

5.5 Distribution and accuracy of question level

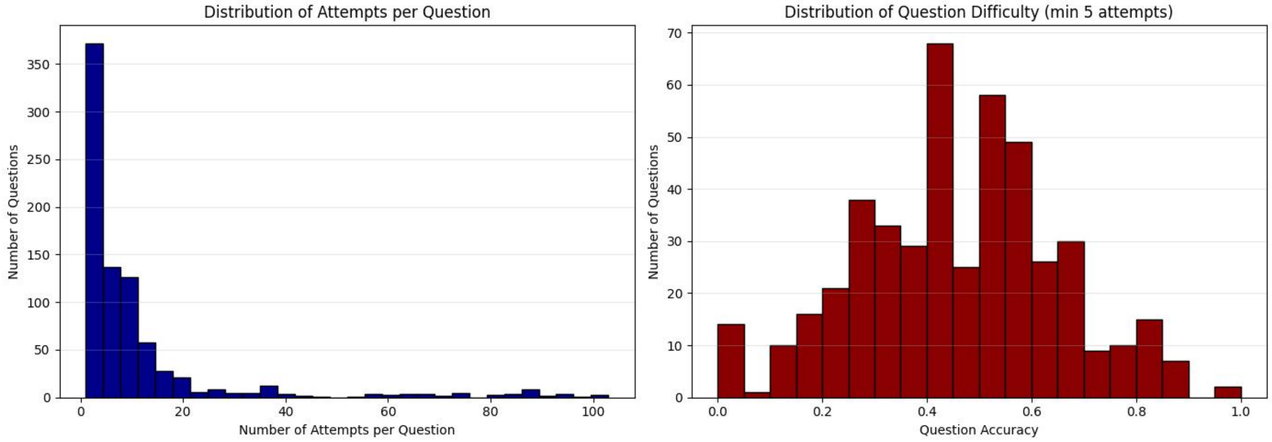






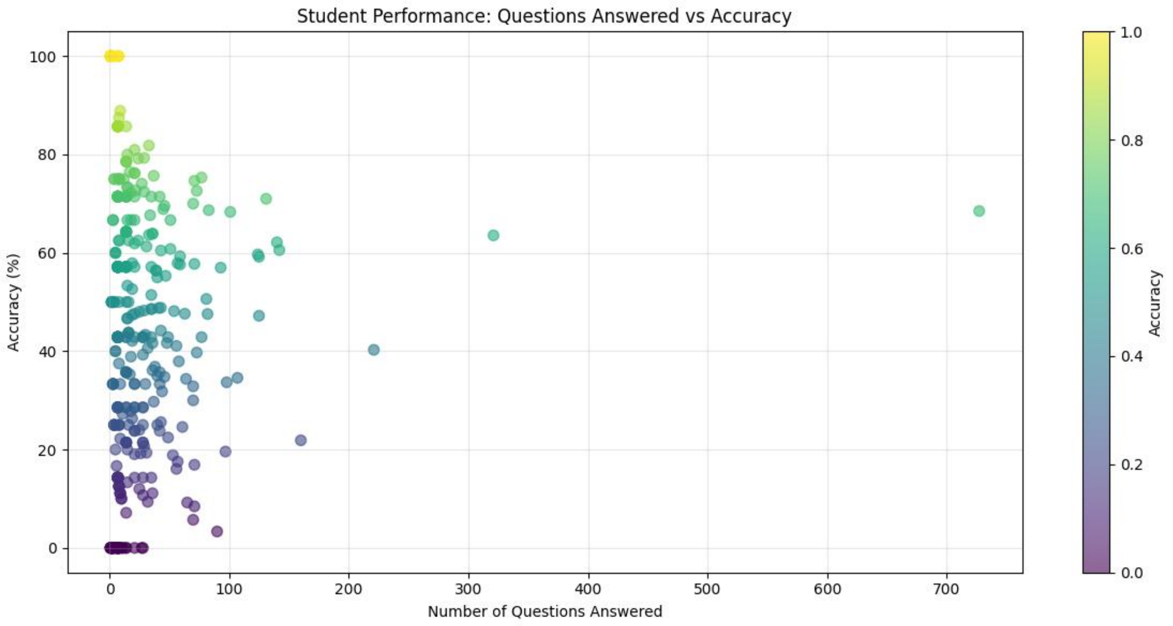
The left graphic shows the distribution of question levels (Basic and Advanced). The number of basic questions is almost 8000, while the number of advanced questions is almost 2000, so there is a prevalence of basic questions. The right graphic shows the accuracy distribution by question level using a box plot; the distribution of advanced question is larger rather than the basic question.

5.6 Distribution of attempts per question and distribution of question difficulty



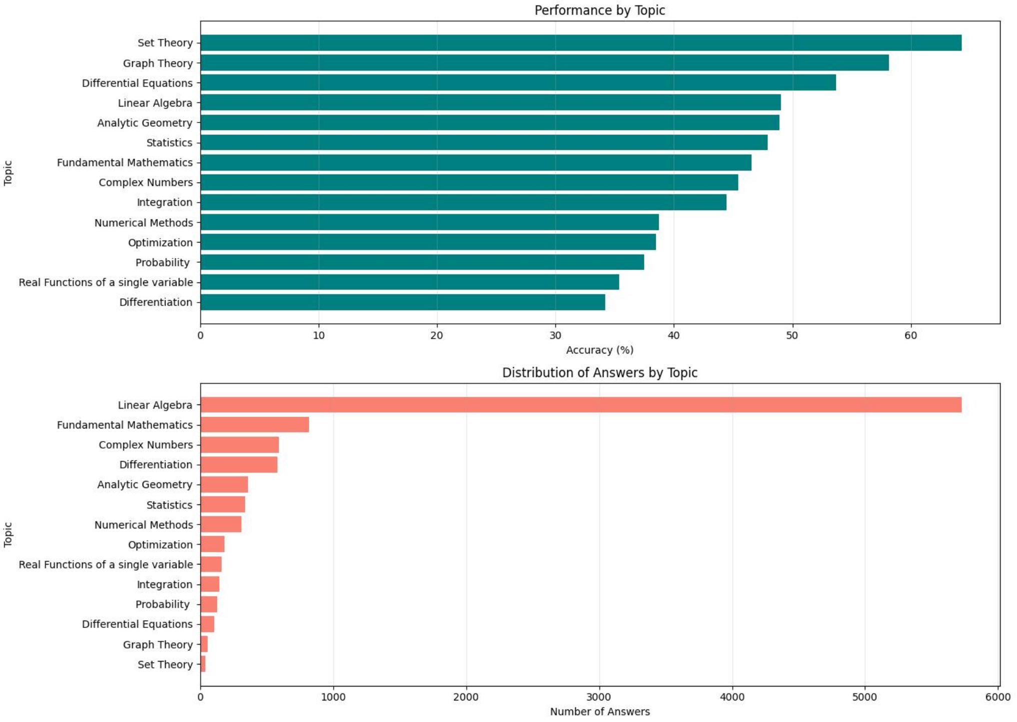
The first histogram shows that more than 350 questions have few number of attempts. The more the number of answer attempts increases, the more the quantity of questions decreases. The red histogram represents the distribution of accuracy of question with a minimum of 5 attempts. We can see that the core of the distribution is in the middle, with a question accuracy between 0.4 and 0.6.

5.7 Student performance: questions answered vs accuracy



In this scatter graphic is represented the student performance evaluated by the number of questions answered and accuracy. We can see that from 0 to 100 questions answered, the accuracy is much variable. The students that have answered more than 100 questions tend to have a decent accuracy.

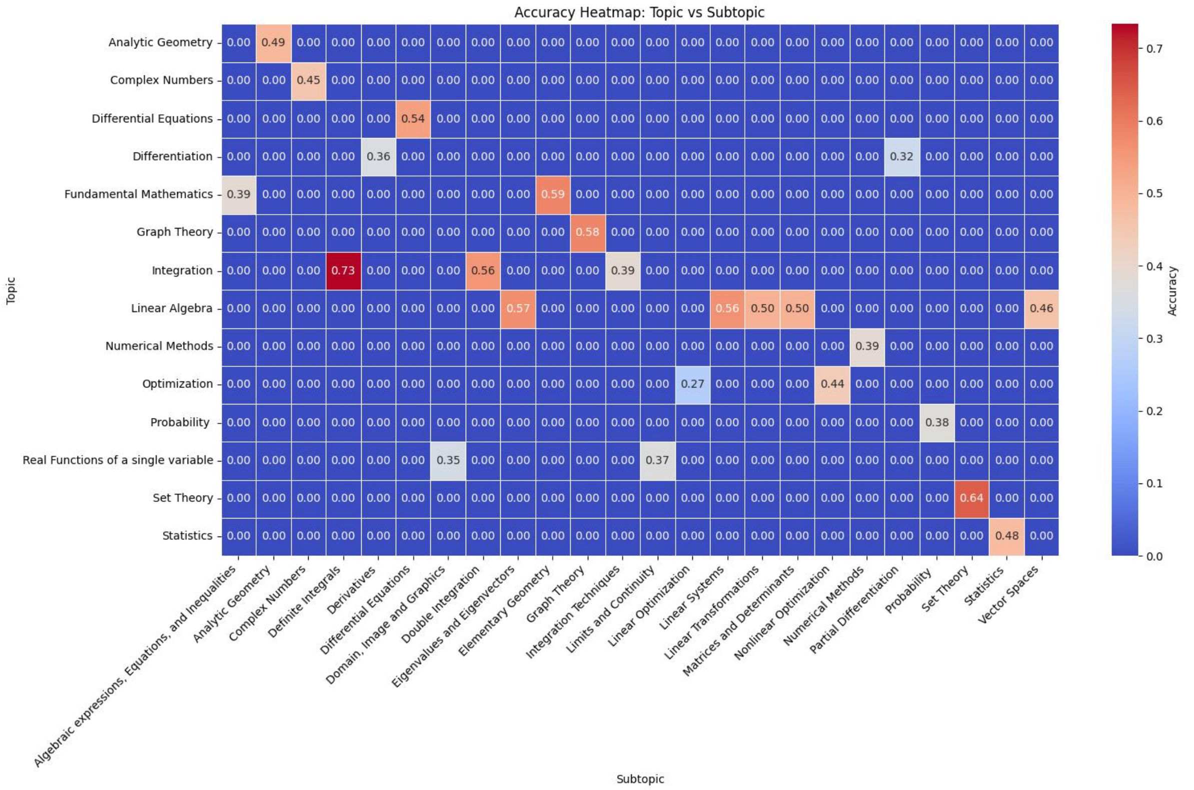
5.8 Performance by Topic



These graphics show the performance and the distribution of answers by topic. We can say that topics like set theory, graph theory or differential equations have the higher accuracy(%)

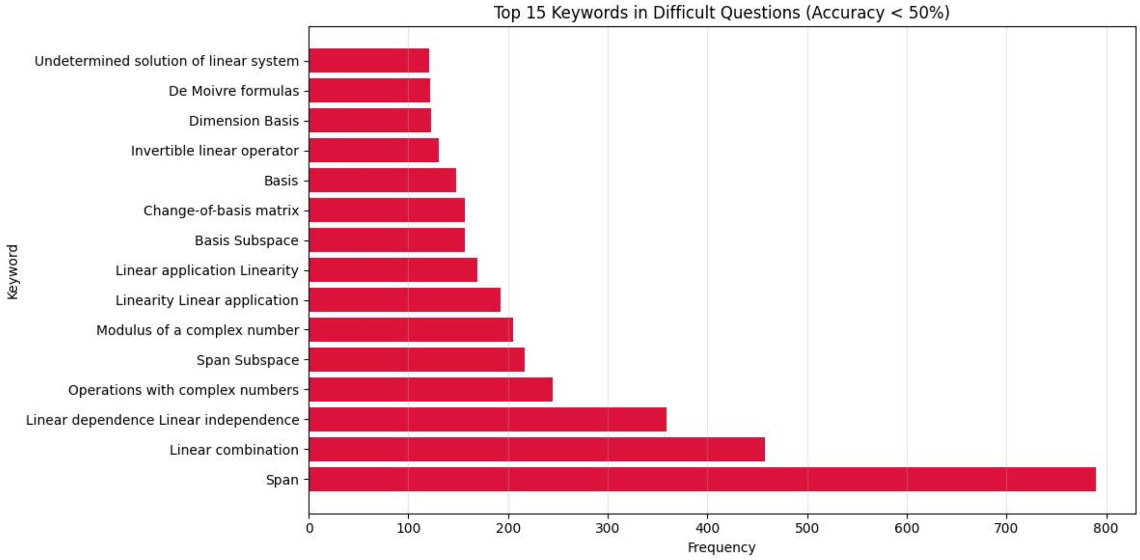
although they have least number of answers. One of the best topics is linear algebra, that has the highest accuracy(%) and the highest number of answers.

5.9 Accuracy heatmap: topic vs subtopic



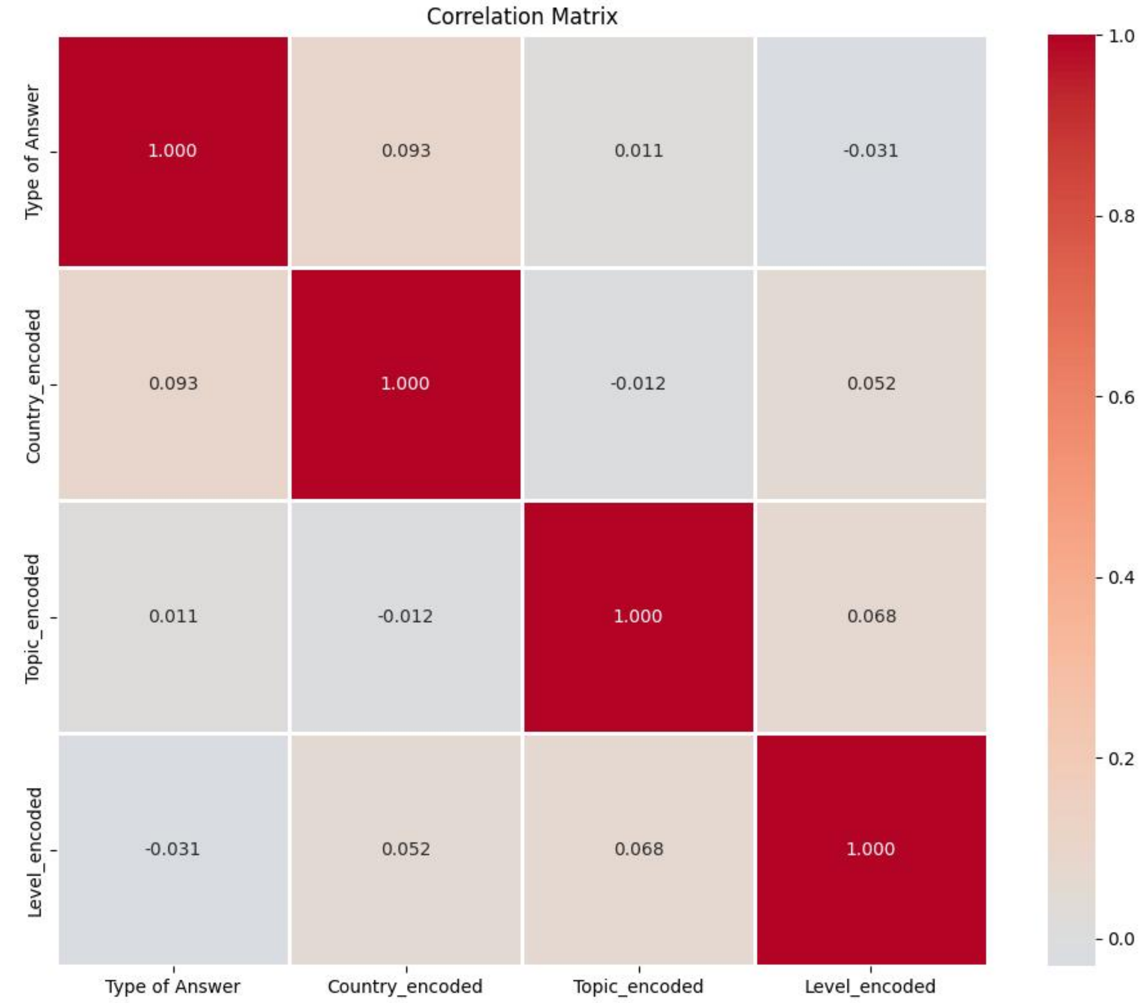
This heatmap shows the correlations between topics and subtopics, ranging from blue (low correlation) to red (high correlation).The highest value is the correlation between integration and definite integrals, which is 0.73, but there are other minor correlations. The majority of correlations is set to 0.00.

5.10 Top 15 keywords in difficult questions (accuracy<50%)



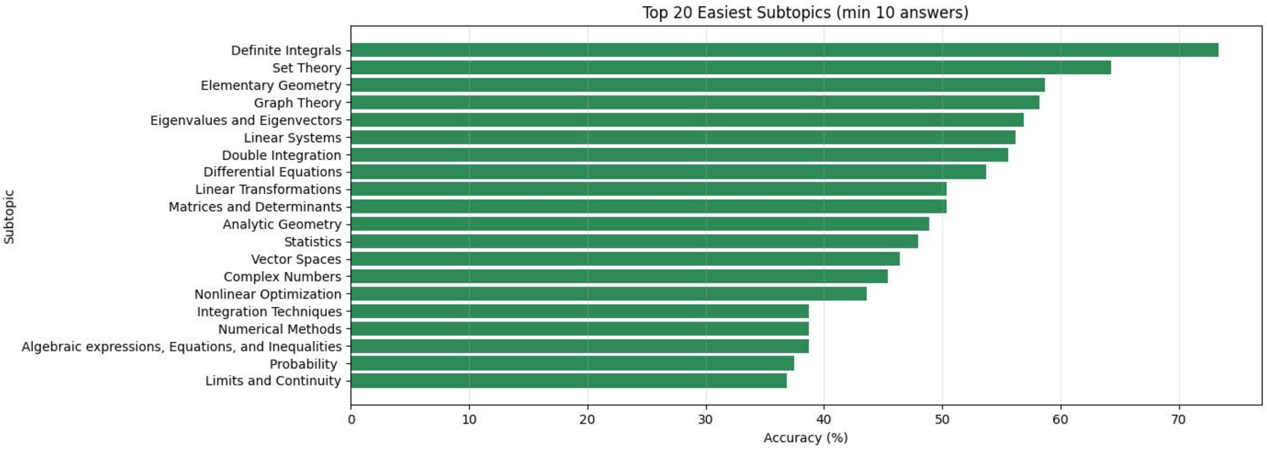
This graphic shows the top 15 keywords that appeared in difficult questions, with an accuracy less than 50%. We might consider that the word "Span" is the most recurring word in difficult questions, with a frequency of 800. This means that the word Span is a common word in math questions, unlike other words that maybe are specific to one or few topics.

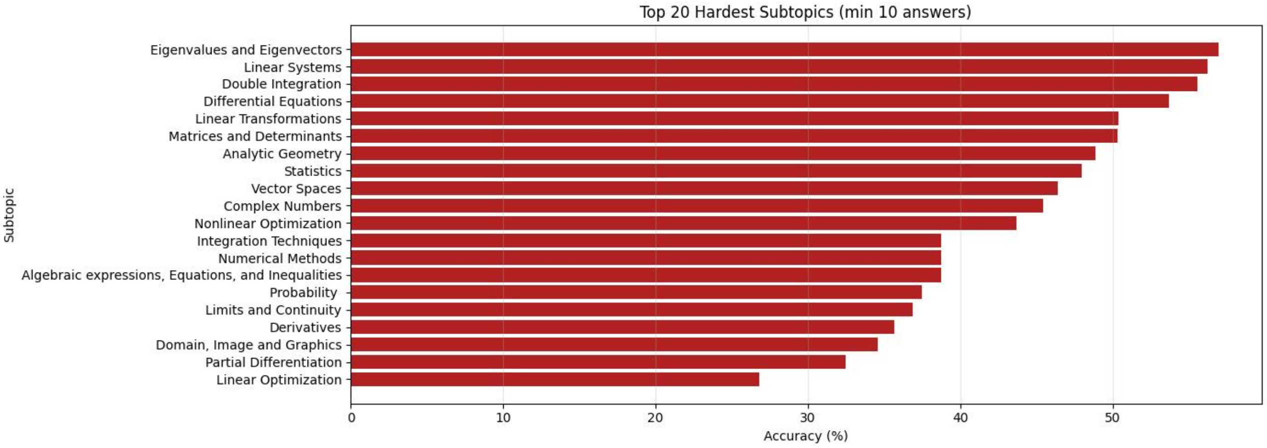
5.11 Correlation Matrix



This correlation matrix helps us to evaluate the relationship of different variables. In this particular case. Correlations like country_encoded and type of answer, with a correlation of 0.093, tells us that the country of the students doesn't have a huge effect on the answers. There are also negative correlations, like level_encoded and type of answers (-0.031).

5.12 Top 20 easiest subtopics and top 20 hardest subtopics (min 10 answers)





These last two graphics represent the top 20 easiest subtopics and the top 20 hardest subtopics, each graphic displays subtopics with a minimum of 10 answers. The easiest subtopic is definite integrals, followed by subtopics like set theory and graph theory, which were also the best topics by performance in the graphic of performance. In the red graphic, subtopics like linear optimization, which is the hardest, doesn't even reach the 30% of accuracy in answers.

6 Model training, evaluation methodology and comparative analysis

In this section we present the modelling phase of the project, whose objective is to predict whether a student answers a question correctly or not, using the features generated during the preprocessing and feature engineering stage. The focus of the modelling process is not only on the predictive performance of the classifiers, but also on the interpretation of the results with respect to the learning behaviour observed in the dataset.

All models were trained on the training split obtained through student–level partitioning. This strategy prevents information leakage between training and test sets and reproduces a realistic scenario, in which the system is evaluated on students that have never been seen during training. In this way, the evaluation measures the ability of the models to generalise to new learners, instead of capturing individual repetition patterns.

6.1 Evaluation metrics and rationale

For each classifier we computed the following metrics on the test set:

- Accuracy: proportion of correctly classified answers;
- Precision: proportion of predicted correct answers that are actually correct;
- Recall: proportion of correct answers that are correctly identified;
- F1–score: harmonic mean between precision and recall;
- ROC–AUC: ability of the model to rank answers by correctness probability.

The joint use of these metrics is particularly relevant in this context. Accuracy alone would not be sufficient, since incorrect answers and correct answers are not evenly distributed across topics, difficulty levels and students. Precision and recall instead allow us to assess how the model behaves in the two error directions: over–estimating success vs. over–estimating failure. Finally, ROC–AUC provides a threshold–independent evaluation, describing whether the model captures a meaningful latent structure in the probability of correctness.

6.2 Models considered

The following models were trained and evaluated:

- Logistic Regression
- Linear Discriminant Analysis (LDA)
- Quadratic Discriminant Analysis (QDA)
- K–Nearest Neighbors (KNN)
- Bernoulli Naive Bayes
- Decision Tree
- Random Forest

These models were chosen because they belong to different modelling families (linear, probabilistic, instance–based and tree–based), allowing us to investigate whether the predictive structure of the dataset is mainly linear or whether significant non–linear interactions emerge.

6.3 Quantitative comparison of results

Table 2 reports the performance of all classifiers on the test set.

Tabella 2: Performance comparison between models on the test set

Model	Accuracy	Precision	Recall	F1–score	ROC–AUC
Logistic Regression	0.635	0.628	0.635	0.629	0.654
Linear Discriminant Analysis	0.617	0.613	0.617	0.614	0.627
Quadratic Discriminant Analysis	0.594	0.589	0.594	0.591	0.586
K–Nearest Neighbors	0.585	0.581	0.585	0.583	0.589
Bernoulli Naive Bayes	0.584	0.516	0.584	0.456	0.518
Random Forest	0.571	0.597	0.571	0.574	0.624
Decision Tree	0.507	0.526	0.507	0.511	0.510
MLP (PyTorch Lightning)	**0.635**	0.574	0.430	0.491	–

6.4 ROC curve analysis

To complement the quantitative comparison reported in Table 2, we analyse the classification behaviour of the models in a threshold–independent way through the ROC (Receiver Operating Characteristic) curves. This representation allows us to evaluate the ranking ability of the classifiers, i.e. their capability to assign higher correctness probability to answers that are actually correct.

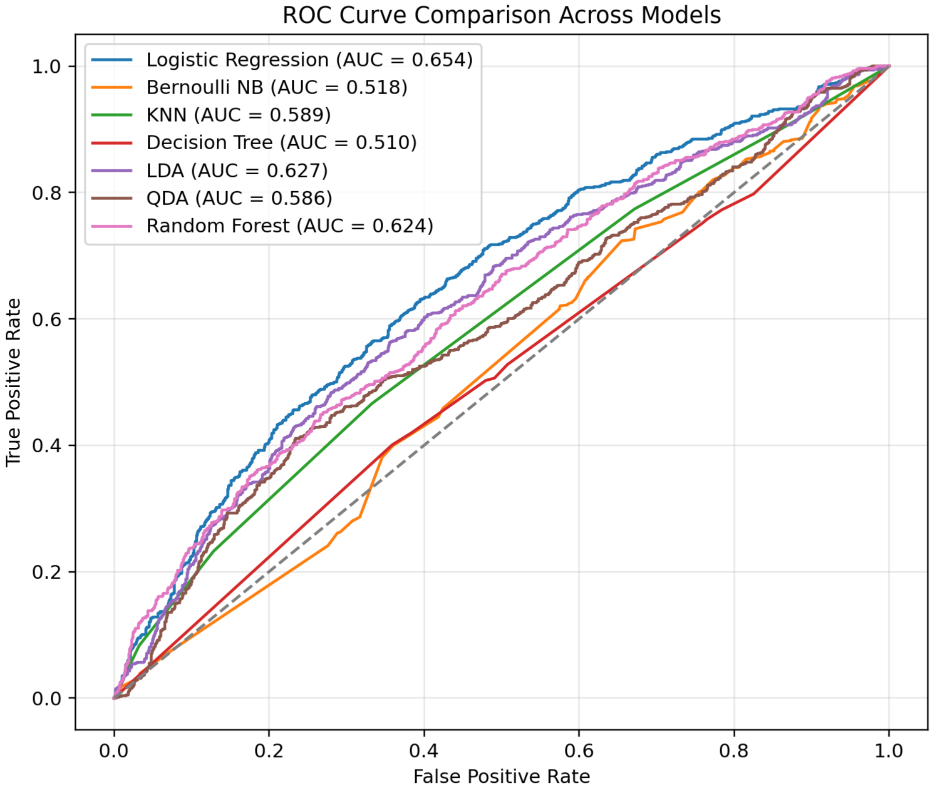


Figura 1: Comparison of ROC curves across all evaluated models.

From Figure 1 we observe that Logistic Regression presents the steepest initial rise and the largest area under the curve (AUC), consistently with the results reported in the metric table.

LDA and Random Forest also achieve a reasonably good separation capability, whereas KNN, QDA and Bernoulli Naive Bayes show curves closer to the diagonal, indicating a weaker discriminative performance. The Decision Tree model is almost aligned with the random baseline, confirming the limitations already highlighted in the quantitative evaluation.

6.5 Interpretation of model behaviour

Logistic Regression is the best performing model both in terms of accuracy and ROC–AUC. This result indicates that the relationship between the engineered features and the probability of answering correctly is largely monotonic and well approximated by a linear decision boundary.

The model does not merely exploit superficial correlations: its performance is consistent with the exploratory analysis, where several features (historical success rate, topic familiarity, empirical question difficulty) showed gradual and interpretable trends in relation to performance. The fact that a linear model performs best suggests that the feature engineering phase successfully captured the most relevant aspects of student behaviour in a structured way.

LDA achieves slightly lower but still competitive performance. Since LDA also assumes linear separation between the classes, this confirms the hypothesis that the predictive signal lies mainly in aggregated and progressive behavioural indicators rather than in local or highly non–linear patterns.

QDA, KNN and Naive Bayes show intermediate performance. In particular:

- QDA benefits from class–specific covariance modelling, but appears more sensitive to noise;
- KNN performs reasonably well, but is affected by sparsity in students with few interactions;
- Naive Bayes suffers from the strong independence assumptions between features, which do not hold in a dataset where variables are intentionally constructed to encode cumulative learning effects.

Tree–based models, and especially the Decision Tree, show lower predictive performance. The Decision Tree tends to overfit local partitions of the feature space and fails to exploit global regularities linked to student experience progression. The Random Forest partially mitigates this effect, but still remains inferior to Logistic Regression. This suggests that the phenomenon under study is not governed by abrupt threshold effects, but rather by smooth behavioural transitions.

6.6 Relation between models and learning dynamics

The superiority of linear models has an important methodological implication. It means that the engineered variables capture underlying learning dynamics that evolve gradually over time:

- students with higher cumulative accuracy tend to maintain higher probability of success;
- familiarity with a topic progressively increases performance on related questions;
- empirical item difficulty interacts with student history in a continuous rather than discrete manner.

Therefore, the model is not only predictive, but also consistent with plausible cognitive mechanisms. The result is aligned with the observations emerging from the EDA: performance is the outcome of the interaction between prior experience, conceptual area and difficulty level, rather than a random fluctuation.

6.7 MLP model: training behaviour and performance analysis

In addition to the classical machine learning models, we also trained a **Multilayer Perceptron (MLP)** implemented using PyTorch Lightning. The objective of this neural model is the same as in the previous experiments, namely predicting whether a student answer is correct or not starting from the engineered feature space. The model consists of two fully connected hidden layers with batch normalisation, ReLU activation and dropout regularisation, followed by a final classification layer with two output units.

Training configuration Training was performed using the same train–test split adopted in the previous experiments, preserving the student–level partition in order to avoid leakage effects and to ensure that the model is evaluated only on students not seen during training. The Adam optimiser was used, with cross–entropy loss as training objective. During training, both training metrics and validation metrics were monitored, including accuracy, precision, recall and F1–score.

Training dynamics From the training log we observe that the model reaches stable training performance quite early. In particular:

- training accuracy increases progressively across epochs, stabilising around values between **0.69 and 0.70**;
- the corresponding F1–score converges to values around **0.69–0.70** as well.

This suggests that the model is able to exploit the structure of the input features and to learn a relatively consistent decision function without exhibiting instability or divergence during optimisation.

Validation behaviour Validation metrics show a different trend:

- validation accuracy oscillates mainly between **0.59 and 0.63**;
- validation F1–score values typically lie between **0.53 and 0.58**;
- the behaviour remains relatively stable across epochs.

This indicates that the model does not suffer from catastrophic overfitting, but at the same time the gap between training and validation metrics remains evident across the whole training process.

Loss behaviour A similar pattern is observed also for the loss values:

- training loss decreases throughout the first epochs;
- validation loss remains higher and shows moderate fluctuations.

This behaviour confirms that the model fits the training distribution more effectively than the test distribution, which is coherent with the fact that the split is performed at student level and therefore test instances correspond to previously unseen learners.

Comparative interpretation From a modelling perspective, this behaviour suggests that the MLP is capable of capturing regularities in the data but does not significantly outperform the linear models evaluated earlier. The validation scores, in particular, remain close to those obtained by Logistic Regression and LDA, confirming that most of the predictive signal is already well described by the aggregated and engineered features, whose relationships are largely monotonic and close to linear.

Generalisation gap The presence of a systematic separation between training and validation metrics can be interpreted as a mild overfitting effect:

- the neural model has enough capacity to adapt more closely to the training students;
- this flexibility does not translate into a substantial generalisation advantage on unseen students;
- this is coherent with the dataset characteristics and with the fact that many features already encode high–level statistical summaries of the student behaviour.

Overall assessment Overall, the MLP confirms the conclusions already emerging from the previous experiments: although neural models can fit the data effectively, the predictive performance is mainly driven by the quality and semantic coherence of the engineered features rather than by the expressive power of the classifier itself. In this sense, the neural network provides an additional validation of the feature design choices adopted in the project, rather than a substantial performance improvement over simpler linear models.

6.8 Strengths and limits of the modelling approach

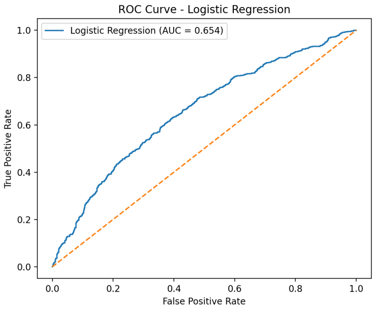
The obtained results are coherent and informative, but some limitations remain:

- the prediction is static and does not explicitly model temporal sequences at interaction level;
- latent student abilities are approximated through engineered aggregates rather than estimated explicitly;
- the current models do not incorporate uncertainty in student ability evolution.

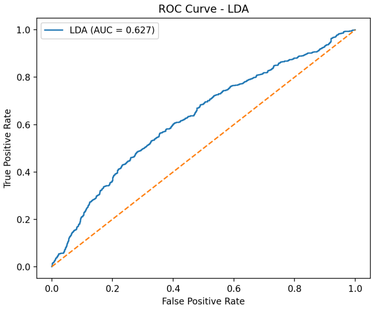
Despite these aspects, the present approach already provides a meaningful and interpretable representation of student performance behaviour, while preserving a strong connection with the evidence highlighted in the exploratory analysis.

Future extensions may include sequential models, but the results obtained here demonstrate that a structured feature engineering pipeline combined with linear modelling constitutes a solid and informative baseline.

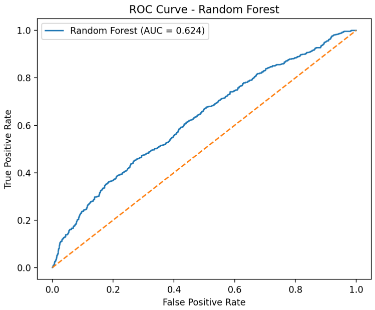
Individual ROC curves



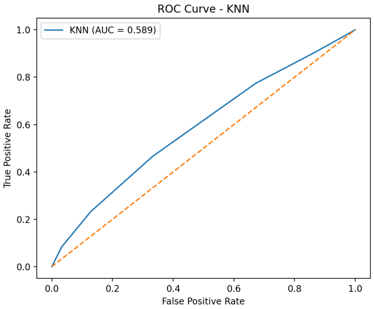
Logistic Regression



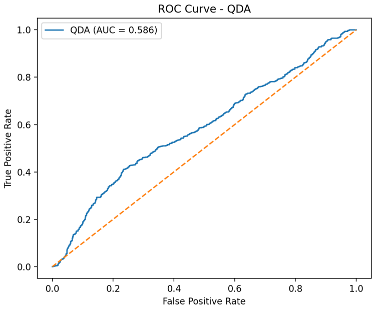
LDA



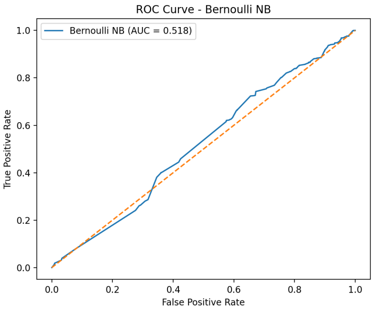
Random Forest



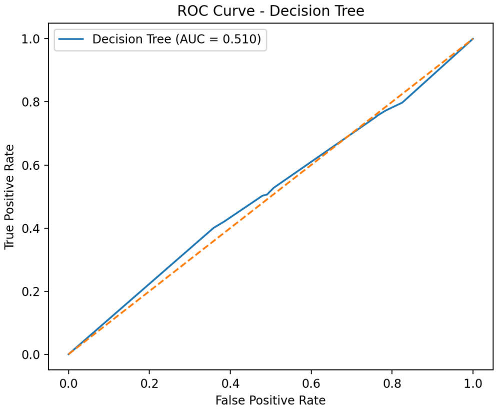
KNN



QDA



Bernoulli NB



Decision Tree