Algorithm Techincal Report

University of Genoa

1 Introduction

Now, questions in users' self-assessment tests on Imath/MathE platforms are randomly chosen among the questions available on the platform's repository. The aim of the Imath project is to use Machine Learning techniques to automatically choose the questions in users' self-assessment tests according to the user's profile (student's background) and past tests' results and data. In this way, each student will be provided with a personal learning path, based on his/her knowledge.

2 Methodology

The approach we propose to choose the questions is the following:

- First question: chosen randomly since there is no information yet on past answers of the current test which can be used in the prediction model.
- From second to last question: using a Random Forest model to predict the probability that the user will correctly answer the next question, for all the possible available questions; then choosing as next question the one which maximizes that probability.

The idea behind this approach is to increase the motivation of the student in studying the subject he's testing himself, cheering him up and leading him to stay longer on the platform, presenting him easy questions (i.e., with high probability that he/she will correctly answer to them). In fact, the platform wants to help students, which have difficulties with math and scientific subjects, to improve their knowledge and to approach more easily this kind of subjects. Moreover, the past surveys of students' opinions on MathE platform have highlighted that for most of the students the questions proposed on the platform were too difficult, suggesting that the approach to propose easy questions could be successful.

The details of the algorithm will be described in the next paragraphs.

2.1 Input data

We use data from different sources as input to the model:

- Data from the informative questionnaire filled in by each student on his/her background. User's grades have been reported in a scale between 0 and 100 to confront grades with different scales. In case of not answered questions or the student does not have performed any math exams yet, we use as input the mode of all the other users' answers. We also added a Boolean variable to indicate if the user effectively performed or not an exam.

- Data from previous questions and tests on the platform, obtained by the data saved by the prototype. For each past question of the current test, we collect: if the answer is correct or not, if it has been skipped or not, the time spent on the question, difficult of the question indicated in the input data, percentage of times the question was correctly answered by the users, number of clicks on the answers. Moreover, we retrieve the percentage of times the current user correctly answered questions in past tests. If it is his/her first test, then the average of all the other users' results is used. We also add a Boolean variable to indicate if the user already performed or not a test.

- Info on the next answer, obtained by data saved by the prototype. For each possible future question, the following quantities are calculated: difficulty of the question indicated by experts, percentage of times the question was correctly answered by the users, average time spent on the question by the users, average number of times the question has been skipped, average number of clicks on the question's answers.

We collected all the initial input features for the models in table 1. These features are first pre-processed: categorical features have been encoded using one-hot encoding, while numerical features have been scaled between (0,1). Then, before giving the features in table 1 in input to the Random Forest models, we performed a features selection stage: only the features with the major influence on the output are selected to feed the models, the others are removed.
In tables 2-3-4-5, there are the final features selected for each Random Forest and used as input by each model.

TOTAL FEATURES
Year of birthday
High school final grade
University year
Love for math (1 to 5)
Student's personal evaluation in math (1 to 5)
Home study hours
Number of passed subjects
Highest grade in the current semester
Average of the positive grades in the previous semester
Average grade in a Mathematics test in the current semester
Highest grade in a Mathematics test in the current semester
Average grade in a Linear Algebra test in the current semester
Highest grade in a Linear Algebra test in the current semester
if he has given any exam or not(0/1)
if he has given any exam in the current semester(0/1)
if he has given any exam in the previous semester or not(0/1)
if he has given any exam in math in the current semester or not(0/1)
if he has given any exam in math in the current semester or not(0/1)
if he has given any exam in linear algebra in the current semester or not(0/1)
if he has given any exam in linear algebra in the current semester or not(0/1)
Percentage of times the current user correctly answered questions in past tests
if he has performed a test on the platform or not(0/1)
If the previous answer(1/2/3/4) is correct or not
If the previous answer(1/2/3/4) has been skipped or not
Time spent on the previous question(1/2/3/4)
Difficulty of the previous question(1/2/3/4) indicated by experts
Number of clicks on the previous question's answers(1/2/3/4)
Statistical difficulty of previous question(1/2/3/4)
If the previous question(1/2/3/4) has been answered or not by any user(0/1)
Difficulty of the possible future question indicated by experts
Statistical difficulty of the possible future question
If the possible future question has been answered(0,1)
Average time spent on the question by the users
If the possible future question has been answered(0,1)
Average number of times the question has been skipped
If the possible future question has been skipped(0,1)
Average number of clicks on the questions' answers
If the possible future question's answers have been clicked(0,1)
Gender
Country
Type of high school
University course
If he/she is a working student or not
Individual or teamwork preferences
Learning methodology
Learning style
Hobby

Table 1: Total Features.

Features of RF question2
Year of birthday
University year
Love for math (1 to 5)
Student's personal evaluation in math (1 to 5)
Home study hours
Number of passed subjects
Average of the positive grades in the previous semester
Average grade in a Mathematics test in the current semester
Highest grade in a Mathematics test in the current semester
Average grade in a Linear Algebra test in the current semester
Highest grade in a Linear Algebra test in the current semester
if he has given any exam or not(0/1)
if he has given any exam in previous semester or not(0/1)
if he has given any exam in math in the current semester or not(0/1)
if he has given any exam in linear algebra in the current semester or not(0/1)
if he has given any exam in linear algebra or not(0/1)
Percentage of times the current user correctly answered questions in past tests
if he has already performed any test on the platform or not(0/1)
If the first answer is correct or not
Time spent on the first question
Difficulty of the first question indicated by experts
Number of clicks on the first answers
Statistical difficulty of first question
Difficulty of the possible future question indicated by experts
Statistical difficulty of the possible future question
Average time spent on the possible future question by the users
Average number of clicks on the possible future questions' answers
Gender
Country
Type of high school
University course
Learning style
Hobby

Table 2: Features table of question2 RF.

Features of RF question3
Student's personal evaluation in math (1 to 5)
Home study hours
Average of the positive grades in the previous semester
Highest grade in a Linear Algebra test in the current semester
If he has given any exam in previous semester or not(0/1)
If he has given any exam in linear algebra or not(0/1)
Percentage of times the current user correctly answered questions in past tests
If he has already performed any test on the platform or not(0/1)
If the second question has been skipped or not
Difficulty of the second question indicated by experts
Statistical difficulty of second question
If the second question has been already answered at least once(0/1)
Difficulty of the possible future question indicated by experts
Statistical difficulty of the possible future question
If the possible future question has been answered at least once(0/1)
Average time spent on the possible future question by the users
Average number of times the question has been skipped
If the possible future question has been skipped at least once(0/1)
Average number of clicks on the possible future questions' answers
If the possible future question's answers has been clicked at least once(0/1)
Gender
Country
Type of high school
University course
Individual or teamwork preferences
Learning style
Hobby

Table 3: Features table of question3 RF.

Features of RF question4
Year of birthday
High school final grade
University year
Number of passed subjects
Highest grade in the current semester
Average grade in a Mathematics test in the current semester
Average grade in a Linear Algebra test in the current semester
if he has given any exam or not(0/1)
if he has given any exam in the current semester or not(0/1)
if he has given any exam in math in the current semester or not(0/1)
if he has given any exam in linear algebra in the current semester or not(0/1)
Percentage of times the current user correctly answered questions in past tests
if he has already performed any test on the platform or not(0/1)
If the first answer is correct or not
Time spent on the first question
Statistical difficulty of first question
Time spent on the second question
Statistical difficulty of second question
If the third answer is correct or not
Time spent on the third question
Difficulty of the possible future question indicated by experts
Statistical difficulty of the possible future question
Average time spent on the possible future question by the users
Average number of times the question has been skipped
If the possible future question has been answered at least once(0/1)
Average number of clicks on the possible future questions' answers
Gender
Country
Type of high school
University course
Learning methodology
Learning style
Hobby

Table 4: Features table of question4 RF.

Features of RF question5
High school final grade
University year
Love for math (1 to 5)
Student's personal evaluation in math (1 to 5)
Home study hours
Number of passed subjects
Highest grade in the current semester
Average of the positive grades in the previous semester
Average grade in a Mathematics test in the current semester
Highest grade in a Mathematics test in the current semester
Average grade in a Linear Algebra test in the current semester
Highest grade in a Linear Algebra test in the current semester
if he has given any exam or not(0/1)
if he has given any exam in the previous semester or not(0/1)
if he has given any exam in math in the current semester or not(0/1)
if he has given any exam in linear algebra in the current semester or not(0/1)
Percentage of times the current user correctly answered questions in past tests
Time spent on the first question
Difficulty of the first question indicated by experts
Number of clicks on the first answers
Statistical difficulty of first question
If the second answer is correct or not
If the second answer has been skipped or not
Time spent on the second question
Difficulty of the second question indicated by experts
Time spent on the third question
Difficulty of the third question indicated by experts
Statistical difficulty of the third question
Time spent on the fourth question
Number of clicks on the fourth answers
Statistical difficulty of fourth question
Difficulty of the possible future question indicated by experts
Statistical difficulty of the possible future question
Average time spent on the possible future question by the users
Average number of times the question has been skipped
If the possible future question has been skipped at least once(0/1)
Average number of clicks on the questions' answers
If the possible future question has been answered at least once(0/1)
Gender
Country
Type of high school
University course
Individual or teamwork preferences
Learning methodology
Learning style
Hobby

Table 5: Features table of question5 RF.

2.2 Output Data

The final output is the index of the next question of the test. The output of the RF, instead, is the probability that the user will correctly answer the next question. This quantity is computed for all the possible tests' questions and the question which maximizes that probability is chosen as next question.

2.3 Algorithm

In order to generate test's questions from 2 to 5 (self-assessment tests consist of 5 questions), a Random Forest is built and trained offline on the available data collected from students from the partner's countries. Then, in the online phase, in order to suggest the student the next question, the trained model is called and used to predict the probability that the student will correctly answer the next question, for each possible question: we have 40 possible math questions among which to choose, but each question cannot be repeated for the same user.

For each question, an optimized RF based prediction model is built. Each RF model differs only for the dimension of past data's window used as input to the model: when defining question 2, the model will have past data only relevant to question 1, when choosing question 3, the model will have data relevant to question 1 and 2 and so on.

The parameters characterising each forest are the following:

- Number of trees in the forest: 900.
- Criterion to choose the best split: entropy.
- Number of features to consider when looking for the best split: square root of the total number of features.

Moreover, we performed a features' importance extraction on the training data to see which feature effectively contributes to the output's prediction and we removed the less important features: we ordered the features according to the computed importance' weights and removed the last elements. We optimized the percentage p of features' cancellation, and these are the results for each model:

- Model to choose question=2: p=60%.
- Model to choose question=3: p=55%.
- Model to choose question=4: p=65%.
- Model to choose question=5: p=60%.

3 Results

In the following paragraph, the results obtained for our proposal have been collected. We saved a portion of data from the training phase to test the algorithm. We computed the prediction for 3 different test sets, each one obtained randomly choosing 10% of the total available data. The following metrics have been used to evaluate the algorithm:

- Classification accuracy (the accuracy in predicting if the student will or not correctly answer the next question).
- False positive rate (the percentage of times when the algorithm wrongly predicted that the student would have correctly answered the next question).

- Complete confusion matrix of the classification's results

The above quantities have been computed for all the 3 test sets and collected in the following tables and figures.

	Test Set 1	Test Set 2	Test Set 3
Classification Accuracy	0.85	0.76	0.74
False Positive Rate	0.22	0.24	0.35

Table 6: Random Forest to predict second question.

	Test Set 1	Test Set 2	Test Set 3
Classification Accuracy	0.76	0.71	0.68
False Positive Rate	0.2	0.25	0.26

Table 7: Random Forest to predict third question.

	Test Set 1	Test Set 2	Test Set 3
Classification Accuracy	0.78	0.68	0.73
False Positive Rate	0.09	0.15	0.07

Table 8: Random Forest to predict fourth question.

	Test Set 1	Test Set 2	Test Set 3
Classification Accuracy	0.84	0.83	0.73
False Positive Rate	0.11	0.09	0.05

Table 9: Random Forest to predict fifth question.

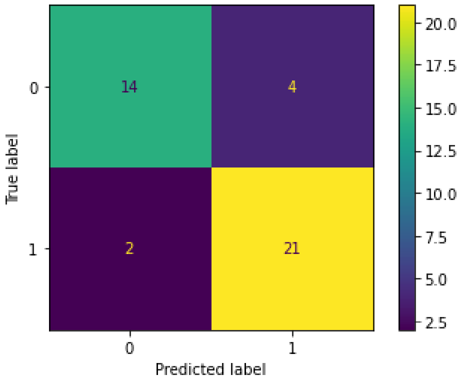

Figure 1: Confusion matrix; Model q=2; Test Set 1.

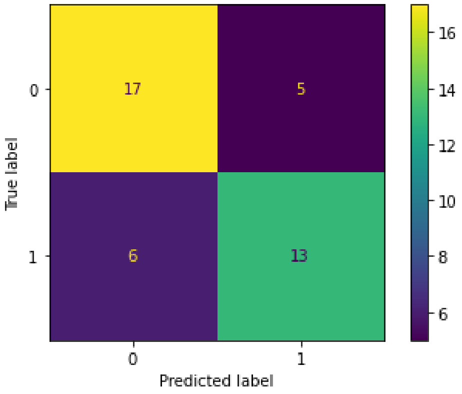

Figure 2: Confusion matrix; Model q=2; Test Set 2.

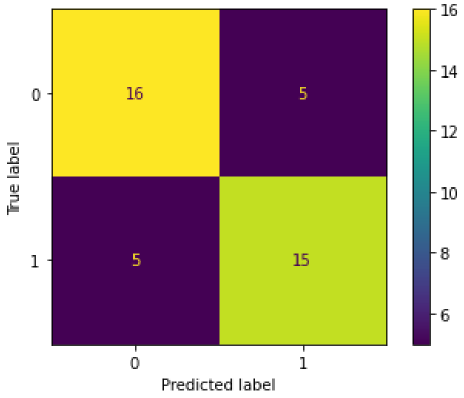

Figure 3: Confusion matrix; Model q=2; Test Set 3.

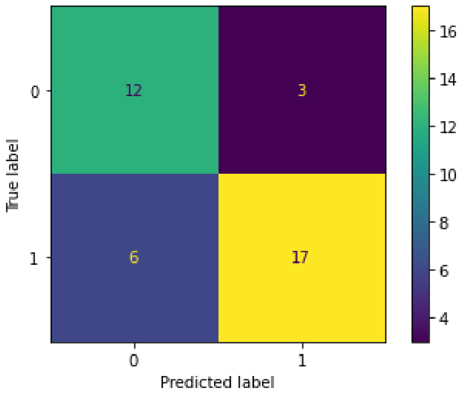

Figure 4: Confusion matrix; Model q=3; Test Set 1.

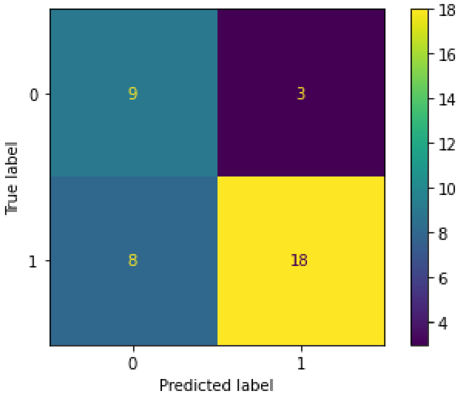

Figure 5: Confusion matrix; Model q=3; Test Set 2.

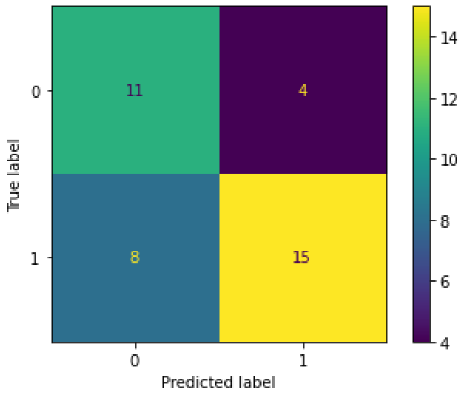

Figure 6: Confusion matrix; Model q=3; Test Set 3.

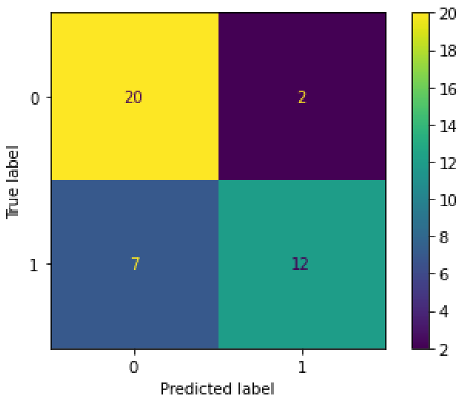

Figure 7: Confusion matrix; Model q=4; Test Set 1.

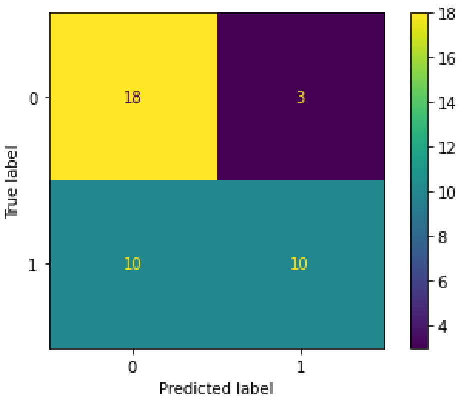

Figure 8: Confusion matrix; Model q=4; Test Set 2.

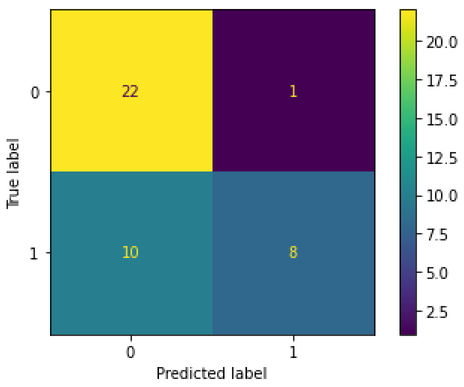

Figure 9: Confusion matrix; Model q=4; Test Set 3.

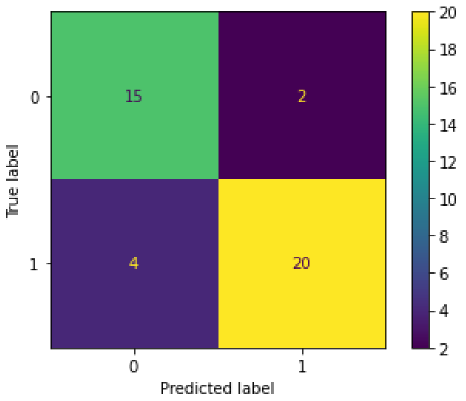

Figure 10: Confusion matrix; Model q=5; Test Set 1.

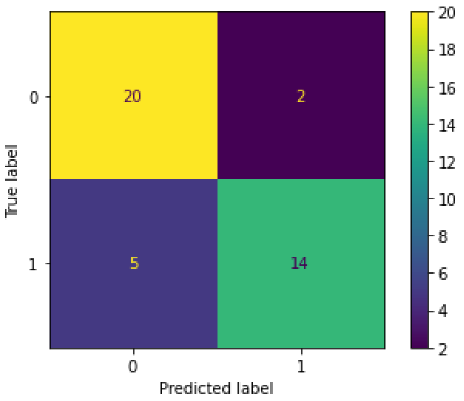

Figure 11: Confusion matrix; Model q=5; Test Set 2.

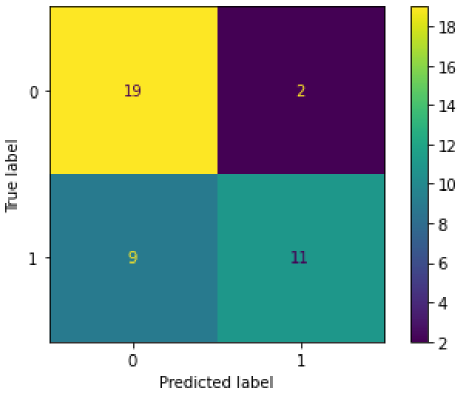

Figure 12: Confusion matrix; Model q=5; Test Set 3.

4 Conclusions

The proposed algorithm, given the amount of data available for training, shows a good accuracy in predicting the probability that the user will correctly answer the next question, and, continuing using the algorithm and collecting more data, the average values (i.e., percentage of correct answers by the user, statistical difficulty of the question, average number of skips on the question,..) will became more precise and the forest will better understand the behaviour of students, improving the quality of the prediction and consequently of the suggestion.