Analysis of Student learning on the EEdi educational platform

Mohammad Ayaan Hashim, Ayshalini Rajahsuresh and Sagarika Coumarane

September 17, 2023

1 Overview

The prevalence of digital resources in education is increasing, giving us the means to measure and adapt more aspects of a student's education. EEdi, an online learning platform, is one such resource that quizzes students with crowd-sourced math diagnostic questions and has them indicate their confidence levels for the same. This data from EEdi was used to analyse the correlation between students' self-reported confidence level and their learning behaviour. In particular, if the self-reported confidence level and their accuracy had any correlation with the subject the test was in, the date of the test, the age of the student and whether their level of income support might have any influence on the student's performance. A logistic regression model was used then to predict whether a student will answer a question correctly based on factors such as their age, gender, whether they are a premium pupil or not, since the output, IsCorrect, is a discrete binary variable taking the values of 0 and 1. We found that there is a positive correlation between a student's self-reported confidence and their learning behaviour.

2 Introduction

Context and motivation Since digital resources are becoming more ubiquitous in education, it is important to understand how to proficiently utilise them to improve a student's learning experience. In order do so, knowing about a student's grasp of concepts and learning behaviours is essential. Assigning online quizzes with diagnostic questions is a common way to do the same.

Diagnostic questions are used to identify misconceptions students may hold about certain topics. Each incorrect answer highlights a specific misconception students could have, thus providing more insights into their understanding of the topic. While answering such questions, students were asked to report their confidence in their answer (on EEdi). This information, when paired with other data (such as demographic, income support, and time of day the question was answered), can be used to draw conclusions about students' learning behaviour. Understanding their learning behaviour would make it easier to create/collate personalised material to efficiently improve their skills.

Previous work Several researchers have investigated the relationship between confidence levels and assessed competence. One study (with almost identical data conditions) concluded that there is a positive correlation between the mean confidence and mean accuracy [2]. Other studies, however, found that there is a negligible correlation between the two[1] [3]. It should be noted that these studies had a similar premise but were conducted in a different subject area for e.g Medicine.

Objectives This data science study explores the correlation between a student's self-reported confidence level and their learning behaviour. We aim to do so by visualising the correlation between confidence level and accuracy as well as the relationship between the available metadata and the confidence level. Additionally, we explore the effects of the time of day, income support, subject group, QuizId, and demographic on student performance.

3 Data

Data provenance The dataset was created by EEdi for the 'NeurIPS 2020 Education Challenge'.[5] Data from the quizzes provided on EEdi from September 2018 to May 2020 was used to draw conclusions in this paper.

As described in the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, the material can be redistributed in any medium or format as long as:

- Appropriate credit is given, a link to the license is provided, and there's an indication of changes made (attribution)
- It's for non commercial purposes
- The modified material isn't distributed

Data description The dataset contains information from 948 diagnostic questions, 4918 students, and 1382727 answers. Each diagnostic question is a math multiple choice question with 4 possible answers.[4] 4 CSV files were provided:

- The general data:
  - QuestionId: the unique identifier for each question.
  - UserId: the unique identifier for each student who answered the question.
  - AnswerId: a unique identifier that corresponds to a specific QuestionID and UserID pair
  - IsCorrect: a binary indication of whether a question was answered correctly (0 if incorrect, 1 if correct).
  - CorrectAnswer: the right option for the given question (number from 1-4).
  - AnswerValue: the option the student submitted for the given question (number from 1-4).
- The subject metadata (where the subjects are organised in a tree structure, with parent nodes and levels to make it easier to group/classify and locate them):
  - SubjectId: the unique identifier for each topic.
  - Name: the name of the topic.
  - ParentId: the unique identifier of the parent node for the given topic.
  - Level: the level at which the subject would be found in the hierarchy.
- The answer metadata:
  - AnswerId: primary key to identify each answer.
  - DateAnswered: time and date the question was answered to the nearest minute.
  - Confidence: the percentage confidence given for the answer (0,25,50,75,100).
  - GroupID: the class in which the question was assigned.
  - QuizID: the quiz which has the question the student answered.
  - SchemeOfWorkID
- The student metadata:
  - UserId: the unique identifier for each student who answered the question.
  - Gender: 0 if unspecified, 1 for female, 2 for male, 3 for other.
  - DateOfBirth: the student's date of birth rounded to the first of the month.
  - PremiumPupil: whether the student is eligible for free meals (pupil premium) or are financial disadvantaged.

Data processing In order to make the data usable, NaN values were dropped where relevant using Pandas. This considerably reduced the size of usable data. Various subsets of the datasets provided were created to make it easier to pick and choose pertinent data and generate new calculated columns. Some datasets were also merged to simplify data access for creating visualisations. External libraries were utilised to train the model for logistic regression.

4 Exploration and analysis

In order to get a clear understanding of how the accuracy of students varies with their confidence, we visualised how the performance of the students varied on fields such as income support, subject(s) entailed by the test, age of the students and date of the exam. We also analysed general patterns in the data, including reasoning about the same.

4.1 Confidence and Accuracy

To find patterns in our primary question (how the students' self-reported confidence while answering a question relates to their accuracy), we calculated the number of submissions made with each confidence interval and the number of correct answers given with that confidence. Then we calculated an 'Accuracy', i.e. $\frac{\text{number of correct answers}}{\text{total number of submissions made}}$.

We then visualised this 'Accuracy (%)' along with the total number of submissions made as a twin plot to make sure that the accuracy was not getting gravely affected by the number of submissions for that confidence level. We did this while making sure that we filtered out all empty values.

From Figure 1 we can see that for all confidence levels, no matter how many submissions were received for that particular confidence, the general trend of the accuracy increased with the confidence level. This led us to infer that a higher confidence in the answer generally led to a higher number of correct answers.

However, there are a few notes that are important to make here:

1. The least amount of submissions made were with a confidence of 25% , at 16981 which is close to half of the confidence level with the second lowest number submissions (0% with 27999 submissions).
2. There also seems to be close to a linear relation among the number of submissions for each confidence levels (apart form the aforementioned confidence level 25%).
3. We can also observe that the confidence levels on the higher end, tended to have far more submissions than those on the lower end.

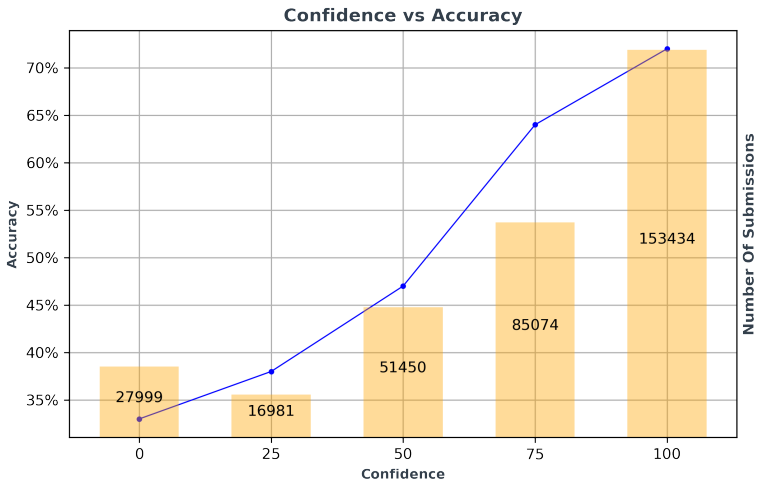

Figure 1: The effect of average confidence on average accuracy.

4.2 Subject and performance

The subject data was organised in a tree structure, where each node had a leaf. The parent node (at level 0) for the entire tree was just maths. This tree had 4 levels to the subjects. Even though it had 4 subject levels to it, we referred just to the second and third subject levels of the tree in our visualisations as we believe that these would be relevant levels when analysing if the subject does have an effect on the students' performance.

To start off with the analysis for the relation between the subject and performance, we visualised the subjects in level 2 along with the 'Accuracy' and 'Average Confidence' giving us Figure 2. This allowed us to check how the accuracy and confidence fared for each of the subjects in this Level (Level 2).

From figure 2, we can see that the difference between the average confidence of the students and the 'Accuracy' varied from about 4% to about 28%. In each of these cases, the 'Accuracy' is lower than the 'Average Confidence'. Therefore it can be concluded that generally, the students are overconfident in their answer. This overconfidence ranges between 4% (in case of the subject 'Indices, Power and Roots') to 28% (in case of the subject 'Decimals').

To further check if a relation contrary to one observed from the level 2 subjects exists, we visualised a similar graph as the one above but in this case, we plotted the 'Accuracy' and 'Confidence' for the subjects in level 3 (Figure 3). We got a similar relation between 'Accuracy' and 'Average Confidence', but in this case instead of all subjects reporting the students as overconfident in their submissions, here, we observed an anomaly, as a subject ('Construct Angle') had its 'Accuracy' higher than the 'Average Confidence'. But this was the only deviation observed between Figure 2 and Figure 3, and can be ruled out as an outlier. Apart from the outlier, the 'Overconfidence' in the level 3 subjects ranged from about 3.5% for 'Squares, Cubes, etc.' to about 48% for 'Enlargement'

Figure 2: The effect of level 2 subjects on performance.

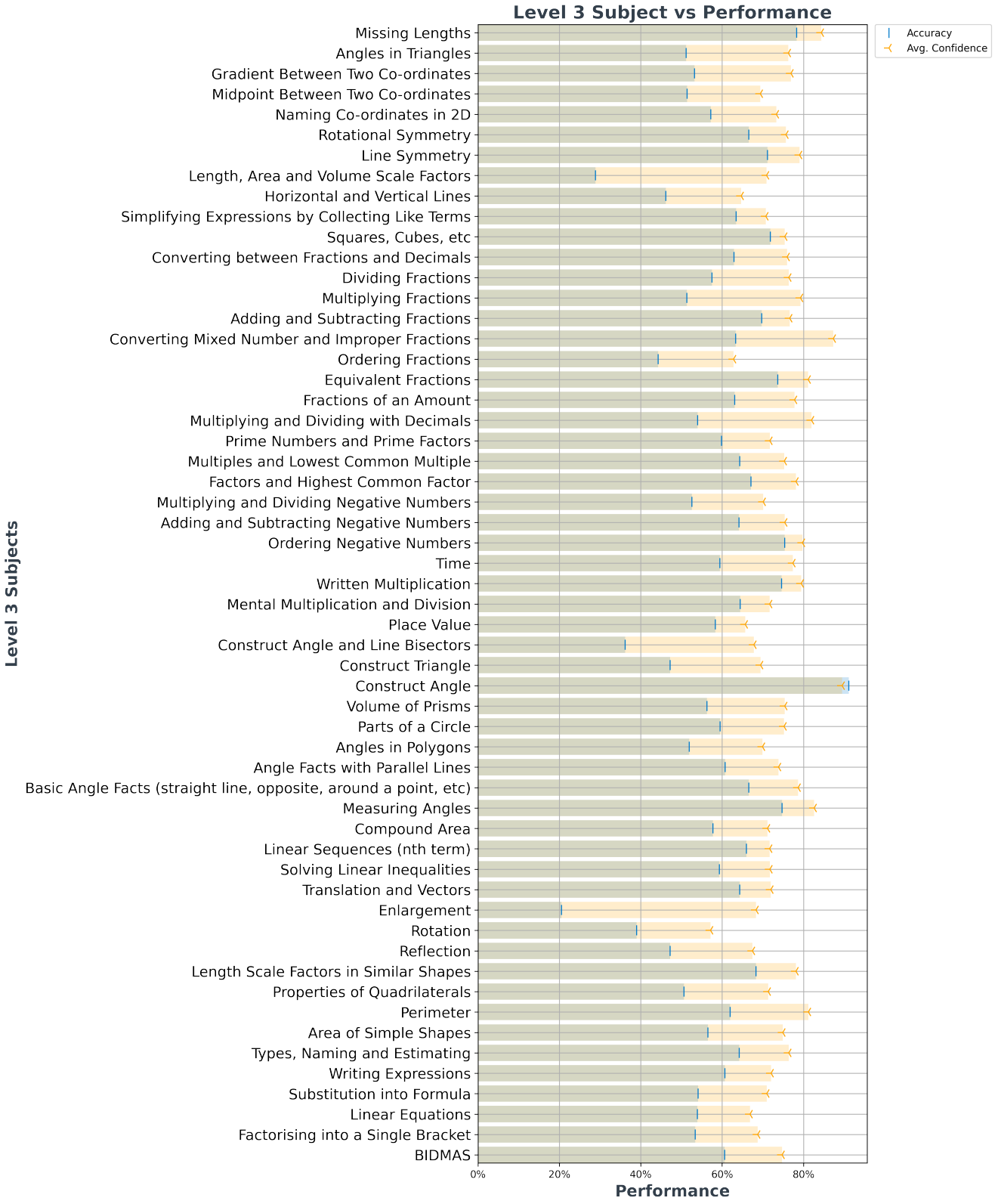

Figure 3: The effect of level 3 subjects on performance.

4.3 Date of quiz submitted and performance

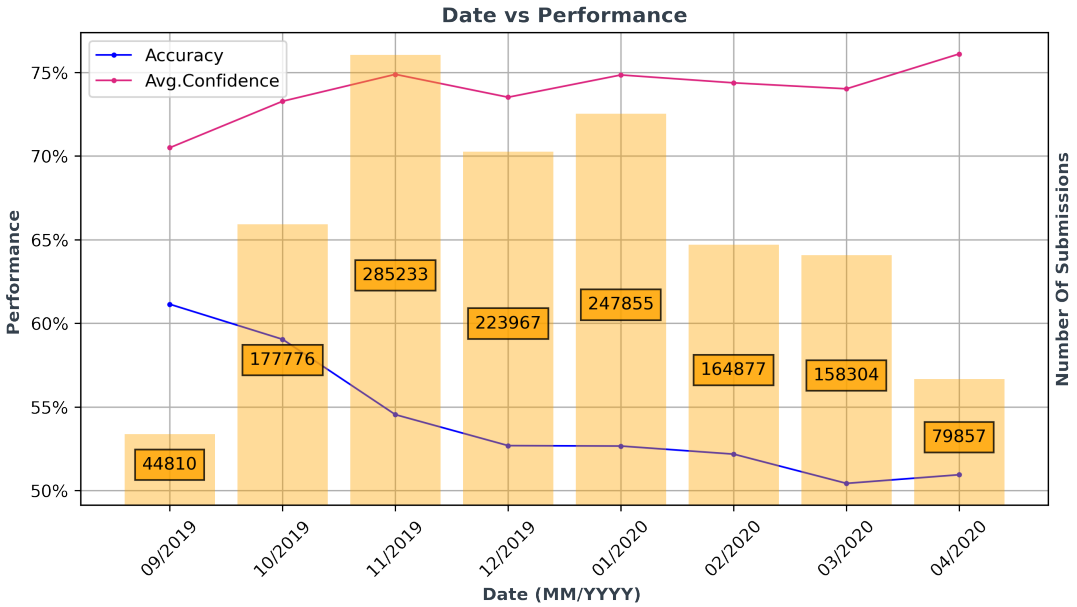

Figure 4: The effect of the date of quiz submission on accuracy and confidence.

For the analysis for the relation between Date of the test and Performance, we plotted the 'Accuracy', 'Average Confidence' and the 'Number of Submissions' for each month year in the files, filtering out the month May of 2020, since there is only one entry for the month (figure. 4). This single entry might be attributed to the raging pandemic, which might also be inferred by the sudden drop in Number of submissions between April and May. But at the same time, the single date entry could also be the last date that the data was collected.

Figure 4 shows that there is an overall decrease in the accuracy whereas there is a general increase in confidence. The number of submissions fluctuated throughout with a peak of 285233 submissions in November 2019. There is no clear correlation between the month year the quiz was taken and performance. March 2020 had the lowest accuracy, while September 2019 had the highest. The average confidence was the lowest in September 2019 while it peaked in March 2020.

4.4 Income support and performance

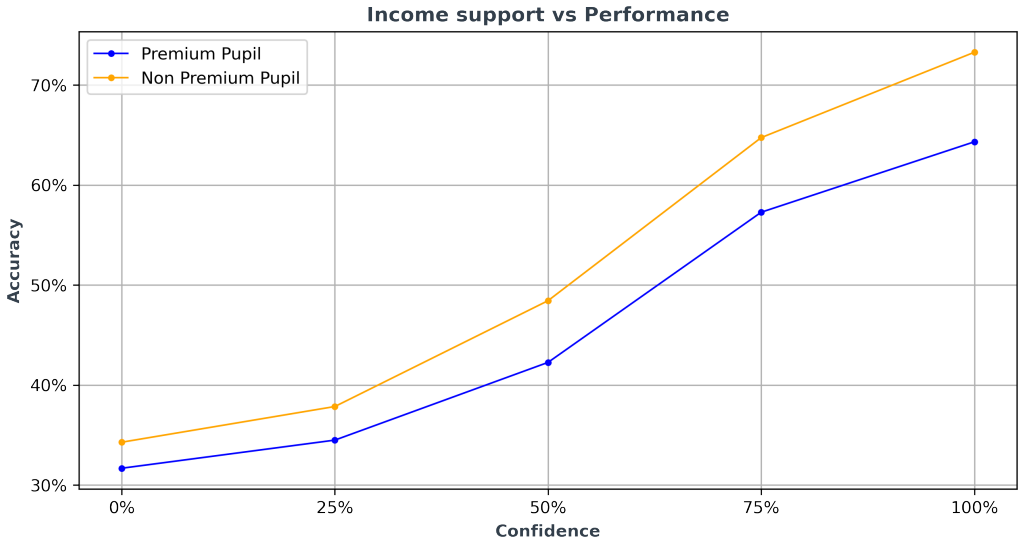

Figure 5: The effect on being a premium pupil on accuracy and confidence.

To investigate whether a student's financial status affects their performance in the test, we decided to look into the trend of performance over each discrete confidence for both premium and non premium pupils. This was calculated by grouping the data by premium/non premium pupil as well as confidence and calculating the average accuracy for each group.

Figure 5 indicates that accuracy increases with confidence regards of whether they are a premium pupil or not. However non premium students obtain more accurate results than premium students across the entire range of self-reported confidence given. This may show that being financially disadvantaged hinders a student's performance during these quizzes.

4.5 Age and performance

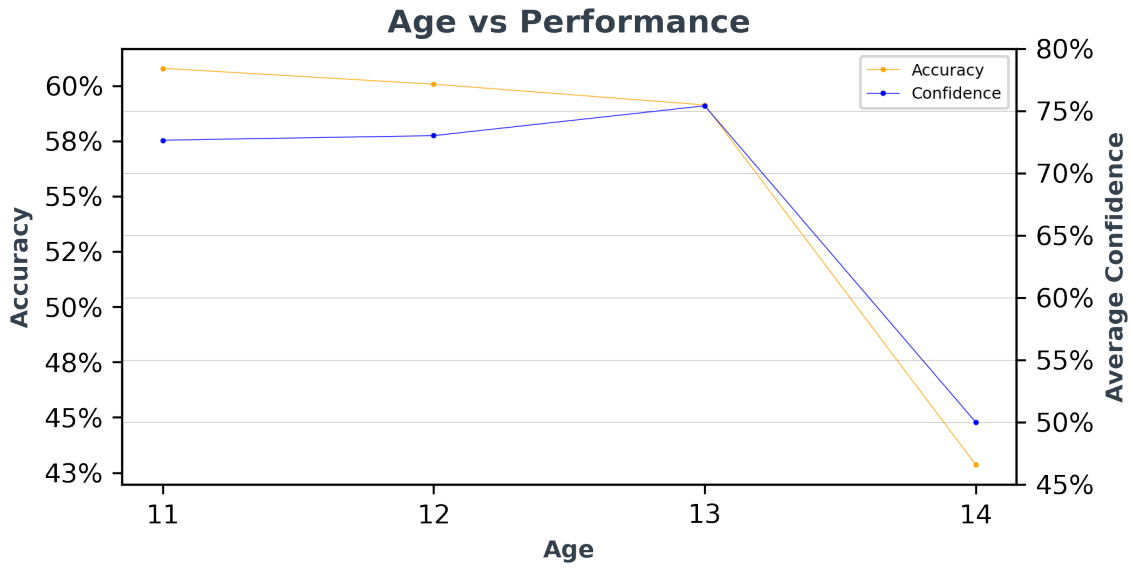

Figure 6: The effect of age on accuracy and confidence.

To investigate how the age of the student might affect the Performance in the test, we plotted the 'Accuracy', 'Average Confidence' with Age (figure.6)

From Figure 6, we can see that both the 'Accuracy' and 'Average Confidence' for ages 11 to 13 are quite similar. But we observe a massive drop in both for 14 year olds, this might be caused because there are only 7 entries for that age. This disparity in the number of submissions for 14 year olds might suggest the fact that the tests(or quizzes) are not appealing to older children. This can be reasoned because the number of submissions declines for older children (82833 entries for 11 year, 78686 for 12 year olds, 14787 for 13 year olds and finally, 7 for 14 year olds).

4.6 Logistic Regression

We decided to use a logistic regression to model whether a student will answer a question correctly or incorrectly based on several factors. The 3 predictors were:

- x_1 - Age.
- x_2 - Gender.
- x_3 - Premium Pupil.

The logistic regression model has been trained using 85% of the available data and tested using the remaining 15% of the data. Figure 7 shows that all p-values are zero which proves that there is strong evidence against the null hypothesis and therefore may be rejected. Thus, it indicates that the predictors mentioned above do affect whether the student gets a question correct or incorrect. The coefficient associated with gender is higher than age which shows that whether the student gets the correct answer is more dependent on gender than age. The coefficient associated with premium pupil is negative so therefore if a student is a premium pupil then they are less likely to get the question correct.

Logit Regression Results

Dep. Variable:	IsCorrect	No. Observations:	176313
Model:	Logit	Df Residuals:	176310
Method:	MLE	Df Model:	2
Date:	Mon, 11 Apr 2022	Pseudo R-squ.:	0.005111
Time:	12:05:51	Log-Likelihood:	-1.1782e+05
converged:	True	LL-Null:	-1.1842e+05
Covariance Type:	nonrobust	LLR p-value:	1.342e-263

	coef	std err	z	P>|z|	[0.025	0.975]
Age	0.0309	0.001	23.139	0.000	0.028	0.034
Gender	0.0997	0.010	10.218	0.000	0.081	0.119
PremiumPupil	-0.3883	0.011	-33.847	0.000	-0.411	-0.366

Figure 7: Summary of logistic regression for predicting the whether a student get a question correct.

5 Discussion and conclusions

Summary of findings Overall, there seems to be a positive correlation a student's confidence levels and their accuracy. Students tend to get more accurate answers when they have given themselves a higher self-reported confidence. There are several other factors that also affect a student's accuracy such as;

age, gender, whether they are a premium student or not and the topic the question is about. Showing how subject affects performance indicates clearly that students are likely to be over confident which their answers.

Evaluation of own work: strengths and limitations One of the strengths of our study is that we have included data that has been collected over a significant period of time which meant that we were able to plot and analyse trends over time such as how the date of when the quiz was taken affects the performance. Moreover, the metadata provided was thorough, thus making it easier to see the larger picture.

We've also developed a predictive model, which makes it easier to understand which factors affect learning behaviour and confidence levels and by how much they affect the results. This has allowed us to quantify and test our hypotheses about certain factors and their effects on a student's accuracy.
There were, however, several missing fields of data which, when removed, cut down the usable size of data to $\approx$25% of the original data.

In the original challenge document[4], the quality of each question was gauged by each student as well as experts. This data unfortunately wasn't available in the datasets provided to us. It is an important field which can be an indicator of a student's learning behaviour, as certain types of questions may be better suited for their learning style. This would have helped us build a more accurate model of learning behaviour.

Comparison with any other related work Our results closely resembled those found by Foster et al. [2]. It didn't however, match the other 2 previously mentioned studies [1] [3]. This may be due to the fact that the field and methods of the studies were different. For example:

- Our study was based in mathematics while the others' studies had to do with medicine
- These studies involved a practical element while ours didn't
- The other studies took the students' confidence levels before the assessment while ours did the same during the assessment

Improvements and extensions One improvement that can be made to the study is to have the same subjects tested over a period of time and note the difference in confidence and learning behaviour. This would provide us with a more accurate indication of any correlation between the two and will help us discern more subtle shifts in learning behaviour (as the data follows a subject and changes over time).

It would also be useful to plot the confidence levels against accuracy and use clustering to identify any significant clusters. This will help us establish a stronger correlation between the two.

Moreover, getting information about the national curriculum/syllabus would also be beneficial to the study. It will allow us to compare the similarity between the quiz questions and the provided curriculum and understand how that variation may cause a difference in learning behaviour, self-reported confidence levels, and accuracy.

References

[1] David J. Brinkman et al. "Self-reported confidence in prescribing skills correlates poorly with assessed competence in fourth-year medical students". In: *The Journal of Clinical Pharmacology* (2015). DOI: https://doi.org/10.1002/jcph.474. eprint: https://accp1.onlinelibrary.wiley.com/doi/pdf/10.1002/jcph.474. URL: https://accp1.onlinelibrary.wiley.com/doi/abs/10.1002/jcph.474.

[2] Colin Foster et al. “School students’ confidence when answering diagnostic questions online - educational studies in mathematics”. In: *SpringerLink* (2021).

[3] Sok Ying Liaw et al. *"Assessment for Simulation Learning Outcomes: A comparison of knowledge and self-reported confidence with observed clinical performance"*. 2011. URL: https://www.sciencedirect.com/science/article/abs/pii/S0260691711002681.

[4] Zichao Wang et al. *"Instructions and Guide for Diagnostic Questions: The NeurIPS 2020 Education Challenge"*. 2020.

[5] Zichao Wang et al. “Diagnostic questions: The neurips 2020 education challenge”. In: *arXiv preprint arXiv:2007.12061* (2020).