# Machine Learning (2150534602) Term Project Report

20182691 Wonseok Do

## Summary:

## Introduction:

Before starting the prediction process, I first looked at the exact contents of the Instructions and Guide for Diagnostic Questions, from what each piece of data means to whether there are specific instructions. First, I looked at the semantics of the primary data (Train Data) and connected it to the metadata by looking at the information in each column. As a result, I assumed that the [QuestionId], [UserId], and [AnswerId] were data associated with the provided metadata. I confirmed that the cases of [IsCorrect], [CorrectAnswer], and [AnswerValue] were the results of the questions the students solved.

Secondly, after checking the other metadata individually, I realized that the provided metadata is not a fully populated dataset. I will have to check these sewing machine data later. Some columns may or may not be useful for deriving results. At first, I will arbitrarily judge the validity of each metadata column based on my personal criteria, and then calculate something likes z-score of the intentionally classified metadata to confirm its validity.

The next issue I thought of was with Rank. The Rank should match the [QuestionId] one-to-one. There is a condition to the ranking, which is as follows; "Rank 1 should correspond to the highest quality question, and so on, in order of decreasing question quality."[1] This means that the quality of the questions should be ranked according to certain criteria. For the ranking criteria, I checked “Results and Insights from Diagnostic Questions”. According to the paper, they use entropy to measure question difficulty and the balance among answer choices. I'm going to adopt the same approach but see if there are other variables that can be applied to it.

## Methods:

I have assumed that all 'Wrong' answers are hard questions.

Confidence in Answer Metadata is the Percentage confidence score given for the answer. 0 means a random guess, 100 means total confidence. However, there were a lot of Missing Values in this data. Even though the values were virtually unusable, I thought the students' Confidence numbers were an indirect indication of the difficulty or clarity of the question. To fill in this value, I thought I'd calculate the probability of appearance based on the distribution of Confidence.

I used a multiple linear regression approach to predict quality. To predict quality, I decided to use KMeans to cluster similar questions together and then assign a score to each cluster based on the characteristics within the cluster. Anyway, since I don't know exactly, I calculated the score by taking the average of the [Confidence] and [IsCorrect] percentages for each cluster. For each data point, I calculated a score based on a combination of [Confidence] and [IsCorrect], which I weighted by the score for the cluster as a whole to get a final quality score.

## Discussion:

When I measured the null data without looking directly at the content of the data, I found that there was one null data in the Test Dataset. However, since the Test Dataset is a private judgment, I decided to ignore it.

Confidence indicates the student's confidence in the question, so it can be an easy question or a clear question. Since no one else can know the source of the confidence, only the person solving the problem, I'll assume that a high-confidence problem is a good problem.

https://github.com/donasensei/QQA

**Conclusion:**

The idea of splitting it into clusters and splitting the order within each cluster seemed plausible at the time. However, I had doubts about the correlation of the values beforehand, and since it was an unsupervised learning method with no right answer, I had to question the whole process. In fact, if I had used a simpler method, I might have gotten higher accuracy, but I wanted to try a new method and didn't have time to tweak the hyperparameters, so I think there's room for improvement for now.

**References:**


1. Wang, Z., Lamb, A., Saveliev, E., Cameron, P., Zaykov, Y., Hernández-Lobato, J. M., … Zhang, C. (2020). Instructions and Guide for Diagnostic Questions: The NeurIPS 2020 Education Challenge (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2007.12061

2. Wang, Zichao, Angus Lamb, Evgeny Saveliev, Pashmina Cameron, Jordan Zaykov, Jose Miguel Hernandez-Lobato, Richard E. Turner, et al. “Results and Insights from Diagnostic Questions: The Neurips 2020 Education Challenge.” PMLR, August 7, 2021. https://proceedings.mlr.press/v133/wang21a.html.


https://github.com/donasensei/QQA