Long-Term Memory Retention of Educational Content

How Machine Learning concepts can be remembered for the rest of our careers with the right practice questions

Ismail Music

Supervisor: Gosia Migut

EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
January 28, 2024

Name of the student: Ismail Music
Final project course: CSE3000 Research Project
Thesis committee: Gosia Migut, Myrthe Tielman

Abstract

To aid the teachings of machine learning (ML), the usage of elaborative interrogative practice questions (EIPQ) is proposed to increase the long-term memory retention of said teaching. Firstly, the existing expectations of students in the current educational landscape are analyzed by taking a look at the undergraduate course present in Delft University of Technology's own Computer Science and Engineering program (CSE2510). Then, relevant theories and techniques for long-term memory retention through practice questions are introduced and applied to CSE2510 content. Finally, an experiment was carried out where roughly half of the participants made use of these newly created EIPQ, while the other half mostly used existing questions, serving as a control group (CQ). The results showed that, compared to the existing practice questions, the use of the newly created EIPQ had a profound impact on the long-term knowledge retention of the learning content. The participants who made use of EIPQ had an average retention ratio of 0.82, compared to the participants who made use of CQ, who had an average retention ratio of 0.57. Therefore, it is suggested that including EIPQ in our current educational model has favorable benefits to the students' knowledge retention of the learned content. A recommendation is made on how to carry out these methods in practice, keeping compatibility with existing learning objectives in mind.

1 Introduction

Machine learning (ML) might be "taking our jobs" in the form of A.I., but subsequently is bringing new opportunities with new challenges, challenges that require adequate knowledge in the field. Adequate knowledge starts with solid foundation, which ML builds at undergraduate teaching levels. Through standardized teaching methods, learning results of high quality can be ensured. Due to the subject's relatively recent spike in popularity, little research has been conducted on its teaching. Machine learning has thus been given a unique position in STEM [1] by being able to innovate how we currently go about its teachings. The question of whether revered STEM teaching methodology is the best fit for machine learning and whether new teaching methodologies should be considered (for undergraduate levels of ML) is at the forefront of this.

The teaching of ML at an undergraduate level has insufficient research conducted on it [2]. Interest in high-school education levels of ML is rising with its teaching methodologies being explored, yet undergraduate curricula can be students' first interaction with the subject. The importance of a topic's first impression cannot be understated [3].

Delft University of Technology has a computer science undergraduate program with a machine learning course, CSE2510. The course is used as a baseline for this research. The first of the learning objectives is partly as follows: "explain the basic concepts and algorithms of machine learning". The course goes over thirteen algorithms and expects the student to be able to explain and code these. It does this over the course of seven weeks, with most algorithms' presence being contained within the week it was first explained. By the end of the course, the students are expected to have retained their memory and understanding of all relevant information about the algorithms.

Retaining a detailed understanding of difficult topics in long-term memory is a difficult task [4]. Teaching methodology in STEM has become aware of a variety of teaching methods that work well to explain these concepts, but with the vast nature of computer science, machine learning specifically, and the overlapping nature of the algorithms you learn, remembering the steps needed for them and the intricacies around this can be an issue in and of itself. This begs the question: is there a way to efficiently teach the students a concept or algorithm while optimizing the way they retain said information?

A technique that has been proven to have some benefits for retention is "**elaborative interrogation**", which is "*generating an explanation for why an explicitly stated fact or concept is true*" [5]. Elaborative interrogation has been proven to increase one's understanding of a topic, but as it stands, measurements of long-term memorization remain insufficient. That being said, even in a subject traditionally seen as an exercise of memorization, anatomy, similar techniques seemed to prove most successful in its teachings [6].

Research varies on the effectiveness of self-generated answers and elaborative interrogation specifically; some studies have shown that failing to generate an answer in such a case will cause a worse performance on the final test than simply reading the correct answer [7]. Proper knowledge of a topic and an explanation of the correct answer thus can negate this somewhat.

This research aims to provide an argument for using elaborative interrogation in machine learning practice questions at undergraduate levels of machine learning. Two methods will be explored for the creation of elaborative interrogative practice questions, EIPQ. Firstly, turning existing ML practice questions into EIPQ. Secondly, creating new EIPQ that make use of elaborative interrogation and real-world machine learning problems. To do so, consider the following:

How can the long-term retention of information given about Machine Learning be increased by using practice questions that implement a form of elaborative interrogation?

To answer this, we consider the defining characteristics of a question that allow for elaborative interrogation. Secondly, we look at how we can turn existing, closed, multiple-choice machine learning questions into elaborative interrogative practice questions. Thirdly, we investigate how to incorporate this elaborative interrogation when creating questions based on real-world machine learning problems. Lastly, we consider the situation in which students can be best presented with these real-world machine learning problems.

2 Related Work

In the following section, the hierarchical model used to explain our learning objects will be introduced. Elaborative interrogation will be further explained, concluding a potential framework. Problem-based learning and problem solving in engineering are then taken a closer look at to draw inspiration from questions about real-world machine learning problems. Finally, we delve into CSE2510 course content and goals, signifying the need for elaborative interrogation and real-world machine learning problems.

2.1 Bloom's Taxonomy

To communicate learning objectives clearly and concisely, Bloom's Taxonomy is commonly used [8]. This consists of a two-dimensional framework that separates Knowledge- and Cognitive Processes. Any learning objective can be represented in these two dimensions through the Taxonomy Table. Depending on its placement, a bloom level is decided. A single learning objective can have several placements in the table. This table can be found in Figure 1.

	The Cognitive Process Dimension					
The Knowledge Dimension	1. ***Remember***	2. ***Understand***	3. ***Apply***	4. ***Analyze***	5. ***Evaluate***	6. ***Create***
A. ***Factual Knowledge***						
B. ***Conceptual Knowledge***						
C. ***Procedural Knowledge***						
D. ***Metacognitive Knowledge***						

Figure 1: the Taxonomy Table

2.2 Elaborative Interrogation

The key theory behind elaborative interrogation is that it enhances learning by supporting the integration of new information with existing prior knowledge [5]. It does so through "why" questions, encouraging the student to elaborate on an explicitly stated fact (new information) with prior knowledge. Most studies have involved self-study. The correlation between prior knowledge and performance through elaborative interrogation shows it benefits students with appropriate prior knowledge. The benefit for lower-knowledge learners is less certain, although in that same study, it was shown that students using elaborative interrogation still relatively outperformed the control group.

The generation effect is the phenomenon of self-generated information being better remembered than when read [9]. In the context of problem-solving tasks, failing to generate an answer can make performance worse [7], which makes sense when you look at low-knowledge concerns. Despite being shown the correct solution after failing to generate the correct answer, these students still performed worse than the students who simply read the solution without ever trying to solve it. For simpler problems, knowledge is retained better when answers to problems are generated by students with lower prior knowledge [10]. Even though generative interrogation showed the opposite, these findings can still be taken into account when thinking of where to implement these types of questions.

A concrete method for creating elaborative interrogative questions is not clearly presented by any of the considered papers. Neither have we found direct usage of this methodology in machine learning or even computer science. It seems there is a gap in knowledge on how to use this technique on more complicated content, with a lack of clarity on what explicitly stated fact the "why" question should apply to [5].

2.3 Problem-based learning

Problem-based learning (PBL) is a well-researched topic. In general, three phases of PBL can be distinguished. The first phase is the pre-discussion, where students work together in small groups on realistic, ill-defined problems. This is a problem that can lead to multiple solutions. Students try to come up with preliminary explanations together based on their prior knowledge and cognitive skills. After the discussion, students collaboratively formulate questions about unclear aspects of the problem. The second phase, the self-study period, has students individually searching for answers in existing literature. The third phase, the reporting phase, has students gather to discuss the found literature and discuss answers to the questions they formulated themselves in phase one. A tutor is present during the first and last phases to ensure all relevant learning objectives have been met [11].

Problem-based learning has somewhat conflicting research results. Previous research has shown that PBL can cause greater enjoyment in learning while maintaining academic performance and, in some cases, increasing it. In other cases, it can cause lower performance on "basic sciences examinations" and even cause a decrease in self-confidence on the topic [12]. While there is enough research that shows the benefits of PBL, the above is a slight indication that it cannot be a replacement for our current ways of teaching and is highly sensitive to the topic at hand.

2.4 Problem solving in Engineering

Students experience a disconnect between engineering classroom problems and work-field problems. Classroom engineering problems are usually generally focused on a single concept and closed-ended, taking away potential engagement for students regarding the topic. When they are open-ended, they tend to be simple problems that do not exhibit the full potential complexity of the topic [13].

One quote given by a student, in particular, exemplified this rigidity in expectations quite well. They stated that the separation of topics spoils the kind of answer you start looking for. Rather than learning how to analyze a problem and feeling prepared for the real-life equivalent of this, you learn how to set your expectations for the examination right and learn based on what you are expected to know.

A systematic approach starts being applied due to this lack of variety in topics, entirely losing part of the problem-solving process. Students have shown to positively respond to applicable problems and feel boosted in their motivation. This underlines the importance of allowing students to feel

engaged with the material by allowing them to think for themselves rather than relying on a rigid structure of expectations.

2.5 TU Delft machine learning (CSE2510)

TU Delft's Computer Science Bachelor's machine learning course, CSE2510, introduces the student to basic machine learning concepts, explains supervised / unsupervised machine learning, and explains the different tasks associated with these types of machine learning. The required understanding of these concepts and algorithms is described through course-level and module-level learning objectives. A course-level objective is expected to be achieved after completion of the entire course, and a module-level learning objective is expected to be achieved after completion of a specific module. An example of a module-level learning objective would be: "After practicing with the concepts of this week, you are able to: explain the basic ideas of machine learning and why and when it can be used." Such learning objectives need to follow certain criteria, easily identified through SMART [14].

CSE2510 introduces new algorithms on a weekly or lecture basis, with all the exercises regarding these algorithms being strictly confined to those weeks. These exercises include online practice questions, coding exercises, and in-lecture questions. With this categorized and segregated approach to the content, students only need to interact with a particular algorithm in the week it was introduced. Much like it was expanded upon in 2.3, this causes a systematic approach per concept, showing a gap in how to consistently keep the student engaged with the material.

3 Methodology

To answer the research question, an empirical, quantitative experiment was conducted. Firstly, the steps taken for said experiment can be found. After, the separate elements of this experiment are elaborated upon. Finally, the sample size determination is explained.

3.1 Experiment

The experiment was conducted over a three-day period. It aimed for 42 participants and had 30 participants complete the experiment.

On day one, the participants were asked about their previous experience with machine learning. Preferably, no participant had any experience in the field, to make sure all their knowledge came from the same source, namely an educational video. This educational video was shown to the participants next. It encompasses basic machine learning concepts and two machine learning algorithms in simple terms. After this, the participants were tasked with answering one of two sets of practice questions and were assigned to an experimental or control group at random:

- (experimental) group 1: elaborative interrogation practice questions (EIPQ)
- (control) group 2: mostly closed, multiple-choice practice questions (CQ)

Group 1 had 16 participants, while group 2 had 14 participants. All participants had the option to look at the answer sheet after answering their assigned set of practice questions, after which they were instructed not to study the material any further. These practice questions can be found in Figure 2 and the Microsoft Forms used to conduct this part of the experiment as well as the answer sheet can be found in Appendices B and C.

On day three, all participants were tasked with answering a set of testing questions identical to CQ, accompanied by an additional question to gauge the participant's confidence in their answer.

A two-day interval was used. Such an interval can be commonly found in long-term retention studies [8]. It was also assumed that a computer science student does not always interact with the subject's material after the day of the lecture in preparation for other courses.

3.2 Learning Content

Nearly all learning content was taken from CSE2510 as this research served as a case study for that course specifically. The following learning objectives were kept in mind: "explain the basic ideas of machine learning and why and when it can be used," "evaluate a trained Machine Learning model" and "explain why a training and test dataset is needed".

As this course is the introduction to machine learning for many students, these students were our target audience. A larger portion of the content contains math, which these students, at this point, had not been taught yet. As the experiment involved taking existing questions and turning them into elaborative interrogative questions, and with the target audience being computer science students who had not taken any machine learning courses yet, only introductory concepts involving minimal math were chosen. The following basic ML concepts are thus to be explained: datasets, models, objects, features, feature representation, vector/feature space, dataset splits (train set/test set), a model's fit, supervised machine learning, and classification. All of the explanations of these concepts were taken from CSE2510 lectures 1.1 [15], 1.2 [16], 3.1 [17], and 3.2 [18].

Two algorithms were explained in a shallow manner, meaning all math was explained intuitively, and no coding application of the algorithms was discussed. The two algorithms were k-nearest neighbors and parzen density estimation when used as classifiers. The explanations for these algorithms were taken from CSE2510 lecture 3.1. For this experiment, the following learning objective was developed, keeping in mind SMART [14] and Bloom's Taxonomy [8]: "Following the instructional video, participants will be able to interpret a Machine Learning classification problem and be able to differentiate their choice between k-NN and Parzen accordingly". Several learning objectives from CSE2510 were taken as inspiration for this, under which "differentiate between different types of dimensionality reduction techniques". All learning objectives either used in the experiment directly or used as inspiration for the experiment can be found in Appendix A.

3.3 Video design

The video is just over seven minutes long. Aiming for a large sample size, the video was kept as short as possible to be mindful of the participant's time. It features an instructor standing on the side of a gray square where imagery and definitions are shown as the topics come by. The video starts with an anecdote about artificial intelligence and proceeds to introduce the listener to machine learning. It then goes over the basic ideas of machine learning and why and when they can be used. Then the two algorithms are explained intuitively, meaning void of any math, and for both algorithms, a real-world example is given. Finally, another real-world problem is given, with the intent of showing how to intuitively decide what algorithm to apply.

The video's set-up was directly inspired by two example videos: a TU Delft pre-lecture video from the Probability Theory and Statistics course [19], and a Socratica abstract math instructional video [20]. The TU Delft video features explanatory imagery and a relatively anecdotal approach to explaining the concepts. The Socratica abstract math instructional video dictated the structure and pacing of the video. Socratica's instructor was also taken as inspiration for the presentation style, meaning mannerisms, sentence structure, and intonation. The educational video used in the experiment can be found in Appendix D, along with the two videos used as inspiration.

3.4 Question design

The key to elaborative interrogation was determined to be letting the participant generate an explanation for an explicitly stated fact. It was also deemed important to facilitate a connection between new information and known information.

Two methods of question creation were developed for this experiment by incorporating these defining characteristics of elaborative interrogation: 1. turning existing closed, multiple-choice questions into questions that use elaborative interrogation. 2. creating new questions that make use of elaborative interrogation and incorporate a real-world example.

The first method went as follows: take any closed, multiple-choice question and put its answer in the question while asking the participant why this is correct. An example of this: "With an increasing number of training samples, overfitting typically becomes more difficult" would become "Why does overfitting typically become less difficult the more training samples we use?"

Instead of choosing the correct answer from a set of answers, the participant generated a reason as to why something was correct. This change did not affect the bloom level. Generating the correct answer to the original question along with an explanation would have tested more factors than just elaborative interrogation; the generation effect would have come into play, providing a possible innate advantage to EIPQ. It has been extensively researched how generating a response will make you remember it better. Despite that, incorporating this could have even possibly served as a bottleneck of the participant's cognitive capabilities [4].

The second method went as follows: take any machine learning problem and ask the user why they would use a particular algorithm. This method had flexibility; it could introduce an answer, but it did not have to. An example would be: "This classification problem involves the identification of handwritten digits. The MNIST dataset is used, which consists of 28x28 pixel grayscale images of handwritten digits (0 through 9)." Then, you could choose between asking: "What algorithm would you here and why" and "Why would you use Parzen here over algorithm x?" [21]

This method combined real-life data problems with elaborative interrogation. Real-life problems have been shown to motivate and facilitate deeper understanding, but the main goal of this research was to examine its effect on long-term memory retention. Two ways to do this were proposed: letting the student explain an answer given about a real-world data problem and letting the student answer themselves and explain further on that. The latter has great benefits for the general memorization of the various algorithms found in CSE2510. Due to the way the course splits up the types of algorithms per week, a student may not interact with a particular algorithm anymore after its initial introduction. Letting the student answer the aforementioned questions allows them to compare the algorithms they have not talked about in that particular week, reminding themselves of these algorithms and their characteristics. It is thus important that, when using these questions, their answer is not always one of the algorithms introduced in the lecture that day.

Eight questions were based on method one, with all eight questions being actual practice questions present in CSE2510. These first eight questions had an existing learning objective from CSE2510 attached to them.

Four questions were based on method two. These newly made last four questions had our previously mentioned learning objective attached to them: "Following the instructional video, participants will be able to interpret a Machine Learning classification problem and be able to differentiate their choice between k-NN and Parzen accordingly." The bloom levels were kept consistent, and the multiple-choice answers for group 2 were developed using certain guidelines [22].

All questions, along with their learning objectives and bloom levels, can be found in Figure 2.

3.5 Sample Size Determination

To determine the required sample size for statistically significant results, a power analysis was conducted through G*Power [23]. The A Priori for the one-tailed t-test with two independent means was used. An α of 0.05 was chosen, a β of 0.2 was chosen, an effect size d of 0.80 was chosen and an allocation ratio of 1 was chosen. These parameters indicate an expectation of a high correlation between the usage of EIPQ and long-term memory retention over EQ. The usage of G*Power can be found in Appendix F.

A sample size n of 42 participants was determined. To aid the acquisition of participants, a raffle was held in which participants were able to win a mini-JBL speaker. On day one, the experiment counted 41 participants, while 30 participants finished the experiment on day three. Group 1 counted 16 participants, group 2 counted 14 participants.

Learning Objective	Question (EQ)	Question (ElPQ)	Bloom Lvl
After practicing with the concepts of this week you are able to explain the basic ideas of machine learning and why and when it can be used	Suppose I gathered data on five different species of Penguin. I measured the height, beak length, weight, feet length, and determined the species for 800 different Penguins. Using this data I would like to build a classifier that predicts the Penguin species. How many features, classes & objects do I have for this problem?	Suppose I gathered data on five different species of Penguin. I measured the height, beak length, weight, feet length, and determined the species for 800 different Penguins. Using this data I would like to build a classifier that predicts the Penguin species. Why do we have 5 classes, 4 features and 800 objects?	B2
After practicing with the concepts of this week you are able to explain the basic ideas of machine learning and why and when it can be used	The arrows in the figure indicate:	Why do the arrows indicate the example's features?	B2
After practicing the topics taught this week, you should be able to evaluate a trained Machine Learning model	The figure shows a fit of a model to some training data. What can be said about the model's fit to the data?	Why can we say the model suffers from underfitting?	B5
After practicing the topics taught this week, you should be able to evaluate a trained Machine Learning model	The figure shows a fit of a model to some training data. What can be said about the model's fit to the data?	Why can we say the model suffers from overfitting?	B5
After practicing the topics taught this week, you should be able to explain why a training and test dataset is needed	With an increasing number of training samples, overfitting typically becomes more difficult	Why does overfitting typically become less difficult the more training samples we use?	B2
After practicing the topics taught this week, you should be able to explain why a training and test dataset is needed	Classification is an unsupervised machine learning problem.	Why is classification a supervised machine learning problem?	B2
After practicing the topics taught this week, you should be able to explain why a training and test dataset is needed	Assume we have trained a classifier h that performs well on the training data. From this, we cannot conclude that h will also perform well on unseen data.	Assume we have a trained classifier h that performs well on the training data. From this, why can we not conclude that h will also perform well on unseen data?	B2
After practicing with the concepts of this week you are able to explain the basic ideas of machine learning and why and when it can be used	An insurance company collects data from its customers. These data include: their income, their age, their monthly premium and the amount of money they have claimed. The company wants to predict whether the customer will stay with the company or leave to another insurance company, based on information of previous customers. They decide to use machine learning to help. What kind of machine learning task is this?	An insurance company collects data from its customers. These data include: their income, their age, their monthly premium and the amount of money they have claimed. The company wants to predict whether the customer will stay with the company or leave to another insurance company, based on information of previous customers. They decide to use machine learning to help. Why is this a classification task?	B2
Following the instructional video, participants will be able to interpret a Machine Learning classification problem and be able to differentiate their choice between k-NN and Parzen accordingly	This classification problem involves the identification of handwritten digits. The MNIST dataset is used, which consists of 28x28 pixel grayscale images of handwritten digits (0 through 9). Could k-NN's classification be used here?	This classification problem involves the identification of handwritten digits. The MNIST dataset is used, which consists of 28x28 pixel grayscale images of handwritten digits (0 through 9). Why would you use k-NN here?	C4
Following the instructional video, participants will be able to interpret a Machine Learning classification problem and be able to differentiate their choice between k-NN and Parzen accordingly	This classification problem involves the identification of handwritten digits. The MNIST dataset is used, which consists of 28x28 pixel grayscale images of handwritten digits (0 through 9). Could Parzen's classification be used here?	This classification problem involves the identification of handwritten digits. The MNIST dataset is used, which consists of 28x28 pixel grayscale images of handwritten digits (0 through 9). Why would you use Parzen here?	C4
Following the instructional video, participants will be able to interpret a Machine Learning classification problem and be able to differentiate their choice between k-NN and Parzen accordingly	A hospital is trying to discover when a patient has an increased chance of a heart disease. They have access to a patient's age, blood pressure, cholesterol levels and whether they have a certain medical condition unrelated to heart diseases. What algorithm would you recommend they use and why?	A hospital is trying to discover when a patient has an increased chance of a heart disease. They have access to a patient's age, blood pressure, cholesterol levels and whether they have a certain medical condition unrelated to heart diseases. What algorithm would you recommend they use and why?	C4
Following the instructional video, participants will be able to interpret a Machine Learning classification problem and be able to differentiate their choice between k-NN and Parzen accordingly	Amazon is trying to see what kind of product it can recommend to a user based on its purchase history. What kind of algorithm can be used here and why?	Amazon is trying to see what kind of product it can recommend to a user based on on its purchase history. What kind of algorithm can be used here and why?	C4

Figure 2: The practice questions per group and their respective learning objective and bloom level

4 Experiment Results

The results from the experiment have been outlined to provide insight into the validity of the question creation methods. Three metrics were used for this: retained correctness ratio, difference in correctly answered questions, and means of confidence levels. The two methods were analyzed separately as well. Prior to no prior knowledge seemed to have insignificant effect. The analysis of the day one and day three scores reveals a significant improvement for the group that made use of elaborative interrogation, whilst the control group scores were stagnant.

In all metrics, each question from day one counted as one point, and on day three the questions counted for $\frac{1}{3}$, $\frac{2}{3}$ or $\frac{3}{3}$ of a point, depending on the level of confidence the participant indicated for said question. While analyzing the data, similar results were found without inclusion of confidence levels affecting the scores, with the biggest difference being in the outliers.

Metric one To gauge the achieved increase of retention by using elaborative interrogation, participants' correctly answered day one questions were compared to the correctness of their day three answer. The data is represented as a ratio of the two scores. For group 2, the CQ group, this resulted in a mean of 0.57 and a standard deviation of 0.16. For group 1, the EIPQ group, this resulted in high increase of the mean with 0.82 and a lower standard deviation with 0.12. This improvement was statically validated by a paired sample t-test, yielding a T-statistic of 4.62 and a P-value of 0.000074. With a Cohen's d value of 1.75 being well above 0.8, this indicates a high impact of correctly understood material being retained in the long-term memory through elaborative interrogation. A box plot showing these ratios can be found in Figure 3.

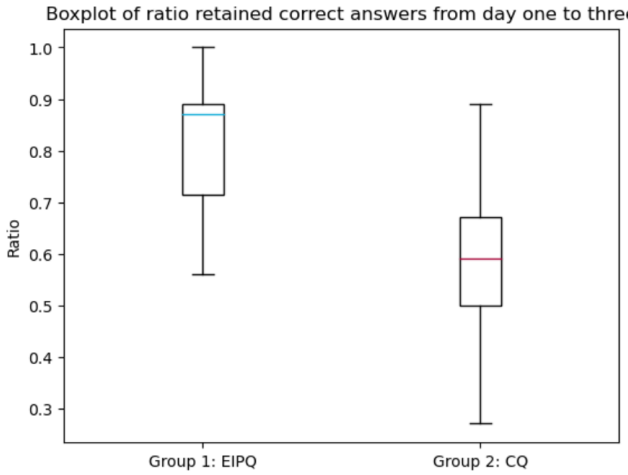

Figure 3: Ratio retained knowledge

Metric two Interestingly, when taking a look at the difference in performance of day one and three when taking incorrectly answered questions on day one into account, participants of group 1 sometimes even showed an increase in correctly answered questions. This did not occur a single time for participants of group 2. For group 2, the mean was 0.70 with a standard deviation of 0.16, indicating a clear decrease in performance. For group 1, this mean was 1.22, indicating a clear increase. The standard deviation for group 1 was quite high with 0.41. Despite this high standard deviation, the T-statistic still showed a value of 4.34 with a P-value of 0.00017 and a Cohen's d of 1.64. This indicates a statistical significance between the two ratios. These two ratios can be found in a box-plot on the left side of Figure 4.

Metric three To show the effect of the confidence levels on these scores, the means of these confidence levels were also compared. Group 2 confidence levels had a mean of 2.01 with a standard deviation of 0.39, while group 1 confidence levels had a mean of 2.52 with a standard deviation of 0.26. With a T-statistic of 4.18 and a P-value of 0.00026 and a Cohen's d of 1.59, this too showed a significant increase by usage of elaborative interrogation. This can be an indication of better retention but also general confidence in knowledge of the material. The confidence level means are shown in a bar chart on the right of Figure 4.

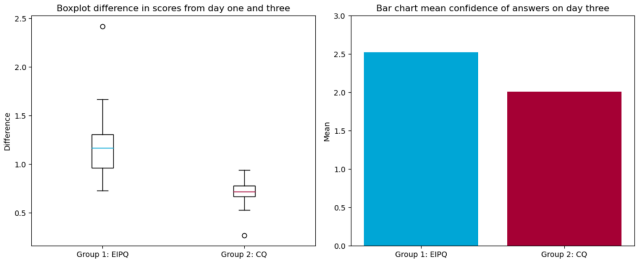

Figure 4: Score difference (left) and confidence level (right)

Method one, turning closed, multiple-choice questions into EIPQ, proved successful in all three of the above-mentioned comparison metrics. For *metric one*, originally answered correctly answered questions still being answered correctly on day three, group 1 had a mean of 0.55 with a standard deviation of 0.14, and group 2 had a mean of 0.45 with a standard deviation of 0.11. The T-statistic was 2.25 with a P-value of 0.03 and a Cohen's d of 0.85, showing a significant statistical improvement in retention. For *metric two*, difference in performance of the entire scores of day one and three, group 1 showed a mean of 1.38 with a standard deviation of 0.53, while group 2 had a mean of 0.75 and a standard deviation of 0.20. This resulted in a T-statistic of 3.98 and a P-value of 0.0004, with a Cohen's d of 1.51. For *metric three*, mean of confidence levels, group 1 had a mean of 2.70 and standard deviation of 0.25 while group 2 had a mean of 2.20 and standard deviation of 0.48. This resulted in a T-statistic of 3.69 and a P-value of 0.00095, with a Cohen's d of 1.4. A box-plot showing the difference in scores of questions 1 to 8 can be found on the left of Figure 5.

Method two, creating new EIPQ based on real-world ML problems, also proved successful in all three of the above-mentioned comparison metrics. For *metric one*, group 1 had a mean of 0.73 with a standard deviation of 0.20, and group 2 had a mean of 0.42 with a standard deviation of 0.30. The T-statistic was 3.56 with a P-value of 0.002 and a Cohen's d of 1.27, showing a significant statistical improvement in retention. For *metric two*, group 1 showed a mean of 1.05 with a standard deviation of 0.48, while group 2 had a mean of 0.66 and a standard deviation of 0.34. This resulted in a T-statistic of 2.47 and a P-value of 0.02, with a Cohen's d of 0.93. For *metric three*, group 1 had a mean of 2.17 and standard deviation of 0.58 while group 2 had a mean of 1.64 and standard deviation of 0.39. This resulted in a T-statistic of 2.81 and a P-value of 0.0090, with a Cohen's d of 1.06. A box-plot showing the difference in scores of questions 9 to 12 can be found on the right of Figure 5.

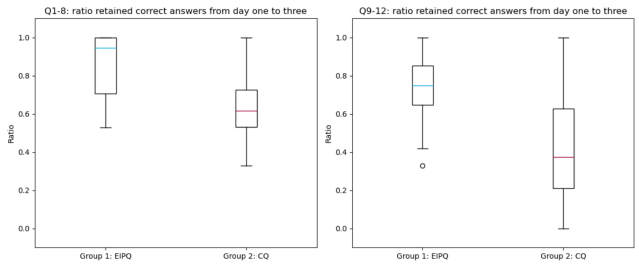

Figure 5: Ratio retained knowledge
method one (left) and method two (right)

Performance scores and confidence scores of participants with prior knowledge were also compared to those without any prior knowledge. Due to difference in scores from day one and three being compared per participant instead of total scores, prior knowledge had no significant effect on any of the metrics utilized above. Prior knowledge actually negatively affected the performance on the above metrics for the elaborative interrogation group, as the only perfect scores achieved on day one in that group were by participants with prior knowledge, which always resulted in a slightly worse score on day three due to confidence being taken into account.

Oral feedback was given by several participants on day one of the experiment. Three participants of group 1 indicated they thought the EIPQ lacked clarity, and this could be seen in the answers given. Some participants provided more detail than needed, and some misinterpreted the question entirely. These were all participants with prior knowledge of the subject. An answer given by a participant in group 1 regarding question one: "To make the model less likely to be negatively affected by outliers in the research material. And so you have different types of measurements since different species can be similar in some ways and different in others". Although this is not the answer the question was looking for, during grading these answers were counted as correct, as to not let misinterpretations influence the score differences. This is advantageous for the CQ group, yet all measurements still pointed towards a clear statistically significant difference in performance favoring the elaborative interrogative group.

Answers to EIPQ were only appointed as correct if they showed understanding of the material, meaning answers were still counted as incorrect if they provided incorrect information about the content being asked.

Five participants in group 1 gave feedback regarding the practice questions seeming difficult. These were all participants with no prior knowledge. They also stated it was due to their inexperience in answering questions like these. When looking at the feedback from the eight participants combined, it seems clarity on the expectations from EIPQ could be improved on. Despite this, all participants who voiced these concerns on specific questions answered correctly on the day three questions with level two or three confidence. Two participants also praised the question quality.

The rest of the oral feedback given regarded the educational video, praising its easy-to-follow nature. This could be an indicator of the lack of influence of prior knowledge on the scores, showing it adequately prepared the participants to answer the practice questions.

5 Discussion

The research done for elaborative interrogation and how this can be incorporated in machine learning courses, CSE2510 specifically, presents significant findings. Our test-based examination of retained knowledge of the content showed a positive correlation between using elaborative interrogative practice questions and the retention of said content.

Three metrics were used to compare using interrogative practice questions to closed, multiple-choice questions. All three showed a statistically significant improvement when using EIPQ. The achieved power for the data indicates that the sample size of 30 was enough for this specific research. In place with what has previously been researched [5], elaborative interrogation indeed serves its purpose for long-term memory retention. This study also mentioned that the more difficult the content, the harder it becomes for elaborative interrogation to be used in practice questions. Both methods one and two show a great example of how this can be done and could serve as a template for any machine learning course making use of practice questions of a similar level of understanding and similarly ranked learning objective according to Bloom's Taxonomy[8]. When we look at the confidence that practicing with elaborative interrogative practice questions brought, it potentially shows a deeper understanding of the material, which would be in line with previous research [5].

Despite these great results, some limitations cannot be understated. The population of the experiment was not ideal; anyone was able to participate, it did not represent the average CSE2510 classroom. There was no confirmation of whether participants looked at the educational video or the provided answers to the questions after answering them on day one. This can be a critical factor in studies like these [7]. Surprisingly, prior knowledge seemed to have no significant effect on the success of EIPQ. Further research would have to confirm the effects of EIPQ on students with prior knowledge of a subject to those without. Of the 16 participants who answered elaborative interrogative practice questions, eight had

prior experience. This could also be a testimony to the educational video, which received praise in the form of oral feedback from the participants, preparing the no-prior experience participants well enough to answer the questions. Oral feedback was also given about the clarity of the EIPQ, stating an increase in perceived difficulty, which is somewhat contrary to previous research saying minimal explanation is required [5]. Despite performance not suffering from this perceived difficulty, an example of such a question being answered in class could prepare the student for when they do it themselves to avoid such struggles. The biggest problem to be considered for this experiment: EIPQ reveal the correct answer to their CQ counterpart in the question itself. Considering there was no way to confirm whether participants examined their answers through the provided answer sheet on day one, this means group 1 participants had an innate advantage by already having seen the correct answer to the questions of day three. This is slightly negated by the results of method two. Here, the correct answer on day three was one of four or five explanations, which group 2 had already seen in their practice questions' answer choices instead. Group 1 still significantly outperformed group 2 on these questions.

Recommendation Based on the study's findings, we recommend the teaching team of CSE2510 and any undergraduate-level machine learning course to consider the use of elaborative interrogation in their practice questions. Furthermore, we also suggest assessing the usage of method two, as the questions created from these do not align entirely with the learning objectives, yet seem great for retention and confidence in knowledge of the learning content.

The defining characteristic of a question that allows for elaborative interrogation is a question that combines prior knowledge with newly found knowledge through the formulation of a "why" question concerning an explicitly stated fact [5].

To turn an existing closed, multiple-choice machine learning question into a question that uses elaborative interrogation, you use the multiple-choice answer and put it into the question as the explicitly stated fact, asking the student why this fact is true. This minimizes the change in bloom level, keeping a consistency to learning objectives and expectations of the course. An answer sheet regarding these questions is recommended.

To create a new practice question around elaborative interrogation and real-world machine learning problems, there are two proposed ways: letting the student explain an answer given about a real-world data problem and letting a student give an answer themselves and explain further on that. The first is recommended if the lecturer expects a detailed answer, omitting the possibility of an unintended cognitive bottleneck [4]. The latter is recommended if a course's learning content is secluded to the weeks it was introduced and if the course has other practice material to make up for the lack of potential detail in the answer.

6 Conclusions and Future Work

To summarize the performed experiment and its results, let us use these to answer *How can the long-term retention of information given about machine learning be increased by using practice questions that implement a form of elaborative interrogation?* Two specific methods of elaborative interrogation were analyzed by conducting an experiment. The first being taking existing closed, multiple-choice questions and turning them into elaborative interrogative practice questions. The second being creating practice questions about real-world machine learning problems and asking the student what algorithm to use or why they would use a certain algorithm through elaborative interrogation. The results of the experiment proved to be significant for both methods, meaning they are valid ways to conduct questions with elaborative interrogation about machine learning, facilitating the retention of knowledge specifically.

This research required a substantial sample size, which influenced the learning content chosen for and used in the experiment. One could argue at the expense of both the reproducibility and validity of the research itself. The simpler the machine learning concepts that are being used and the more shallow its presentation, the easier it will be to remember by default. Despite this, participants with no prior knowledge indicated their struggles with the material and still performed substantially better when making use of elaborative interrogation, indicating a firm practical use of the methodology developed.

For future research, a requirement that should be included is testing the methods on computer science students exclusively. To achieve this, a closer look at their curriculum and thus their knowledge would be required to base the content on. Confidence levels when answering practice questions on day one should also be included.

Subsequently, different time intervals between part one and part two of the experiment can be analyzed. To add, making sure all participants have seen the correct answer at least once on day one is a necessary addition. Furthermore, "self-explanation" could be looked at. A concept similar to elaborative interrogation, yet distinct enough to bring its advantages. One such advantage is an increased performance on concrete problems when with abstract-level practice questions with self-explanation [5]. This could bring a great benefit to method two, the real-world questions. To discover more about this, an experiment similar to ours could be conducted.

In conclusion, this research signifies an underused type of practice question in the form of elaborative interrogation. It explored its possibilities and successfully pinpointed its potential strengths in machine learning. Two methods have been suggested with each its unique strengths relevant to undergraduate education. One of the methods requires minimal changes to the curriculum while the other increases student engagement. This study makes a point of the relevance of good practice material, urging a proactive approach in the creation thereof and the subsequent inclusion of elaborative interrogation.

7 Responsible Research

The TU Delft's FAIR data principles were taken into consideration by allowing transparent data management, methodology documentation, and data sharing to ensure easy repeatability. During the process of this, the Netherlands Code of Conduct for Research Integrity was taken into account. The steps taken (in preparation for) the experiment have been clearly described in section 3 in favor of reproducibility.

The participants were required to give consent to the ethical requirements of the experiment. It ensured the participant was aware of potential data leaks, what that could mean for them, and how their data would be used by the research team itself. The experiment has undergone a review by the Human Research Ethics Committee.

The data was collected through TU Delft's Microsoft Forms server with no log-in required. This allowed the participant to stay entirely anonymous apart from their preferred form of communication. The acquired data was then modified to only include the number of correct answers. This type of data could potentially be tracked back to the participant making it a sensitive data risk, meaning no local storage took place apart from data analysis. All data was analyzed through G*Power and well-regarded Python libraries, minimizing the risk of an incorrect assessment of the significance of the results.

References

[1] re:learn by CcHUB, "Effective Teaching Methods for STEM Education," *Medium*, Feb. 25, 2020. https://medium.com/@relearnNG/effective-teaching-methods-for-stem-education-69f92bb8c6ef

[2] A. J. Ko, "We need to learn how to teach machine learning," *Bits and Behavior*, Aug. 21, 2017. https://medium.com/bits-and-behavior/we-need-to-learn-how-to-teach-machine-learning-acc78bac3ff8

[3] A. Jafar, "The Lasting Impact of a First Impression: An Exercise for the First Day of Class," *Teaching Sociology*, vol. 49, no. 1, pp. 73–84, Dec. 2020. https://doi.org/10.1177/0092055x20966709

[4] R. Duran, A. Zavgorodniaia, and J. Sorva, "Cognitive Load Theory in Computing Education Research: A Review," *ACM Transactions on Computing Education*, vol. 22, no. 4, Apr. 2022. https://doi.org/10.1145/3483843

[5] J. Dunlosky, K. Rawson, and D. Willingham, "Improving Students' Learning With Effective Learning Techniques: Promising Directions From Cognitive and Educational Psychology," *Psychological Science in the Public Interest*, vol. 14, no. 1, pp. 4–58, Jan. 2013. https://doi.org/10.1177/1529100612453266

[6] S. A. Miller, W. Perrotti, D. U. Silverthorn, A. F. Dalley, and K. E. Rarey, "From college to clinic: Reasoning over memorization is key for understanding anatomy," *The Anatomical Record*, vol. 269, no. 2, pp. 69–80, Apr. 2002. https://doi.org/10.1002/ar.10071

[7] J. M. Kizilirmak, B. Wiegmann, and A. Richardson-Klavehn, "Problem Solving as an Encoding Task: A Special Case of the Generation Effect," *The Journal of Problem Solving*, vol. 9, no. 1, Mar. 2016. https://doi.org/10.7771/1932-6246.1182

[8] D. R. Krathwohl, "A Revision of Bloom's Taxonomy: An Overview," *Theory Into Practice*, vol. 41, no. 4, pp. 212–218, 2002. Available: http://www.jstor.org/stable/1477405

[9] S. Bertsch, B. J. Pesta, R. Wiscott, and M. A. McDaniel, "The generation effect: A meta-analytic review," *Memory and Cognition*, vol. 35, no. 2, pp. 201–210, Mar. 2007. https://doi.org/10.3758/bf03193441

[10] B. Rittle-Johnson and A. O. Kmicikewycz, "When generating answers benefits arithmetic skill: The importance of prior knowledge," *Journal of Experimental Child Psychology*, vol. 101, no. 1, pp. 75–81, Sep. 2008. https://doi.org/10.1016/j.jecp.2008.03.001

[11] M. Wijnen, Sofie M. M. Loyens, and L. Schaap, "Experimental evidence of the relative effectiveness of problem-based learning for knowledge acquisition and retention," *Interactive Learning Environments*, 24:8, 1907-1921, 2016. https://doi.org/10.1080/10494820.2015.1060504

[12] M. A. Albanese and S. Mitchell, "Problem-based learning," *Academic Medicine*, vol. 68, no. 1, pp. 52–81, Jan. 1993. https://doi.org/10.1097/00001888-199301000-00012

[13] N. J. McNeill, E. P. Douglas, M. Koro-Ljungberg, D. J. Therriault, and I. Krause, "Undergraduate Students' Beliefs about Engineering Problem Solving," *Journal of Engineering Education*, vol. 105, no. 4, pp. 560–584, Oct. 2016. https://doi.org/10.1002/jee.20150

[14] D. Chatterjee and J. Corral, "How to Write Well-Defined Learning Objectives," *The Journal of Education in Perioperative Medicine : JEPM*, vol. 19, no. 4, Oct. 2017. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5944406/

[15] J. H. Krijthe, M. A. Migut, D. M. J. Tax. (2023). CSE2310 Lecture 1.1: Introduction To Machine Learning [PDF document]. Available: https://brightspace.tudelft.nl/d2l/le/content/595332/viewContent/3517998/View

[16] M. A. Migut, J. H. Krijthe, D. M. J. Tax. (2023). CSE2310 Lecture 1.2: Machine Learning: the basics [PDF document]. Available: https://brightspace.tudelft.nl/d2l/le/content/595332/viewContent/3520844/View

[17] M. A. Migut, J. H. Krijthe, D. M. J. Tax. (2023). CSE2310 Lecture 3.1: Non-parametric density estimation [PDF document]. Available: https://brightspace.tudelft.nl/d2l/le/content/595332/viewContent/3529823/View

[18] M. A. Migut, J. H. Krijthe, D. M. J. Tax. (2023). CSE2310 Lecture 3.1: Non-parametric density estimation [PDF document]. Available: https://brightspace.tudelft.nl/d2l/le/content/595332/viewContent/3529823/View

[19] Delft University of Technology, "Set theory: Union and Intersection - Mathematics - Probability and Statistics - TU Delft," *www.youtube.com*, Jul. 29, 2021. https://www.youtube.com/watch?v=envnkifm9lU

[20] Socratica, "Group Definition (expanded) - Abstract Algebra," *www.youtube.com*, Nov. 07, 2017. https://www.youtube.com/watch?v=g7L_r6zw4-c&list=PLi01XoE8jYoi3SgnnGorR_XOW3IcK-TP6&index=2

[21] T. Penumudy, "A Beginner's Guide to KNN and MNIST Handwritten Digits Recognition using KNN from Scratch," *Analytics Vidhya*, Jan. 29, 2021. https://medium.com/analytics-vidhya/a-beginners-guide-to-knn-and-mnist-handwritten-digits-recognition-using-knn-from-scratch-df6fb982748a

[22] BYU Faculty Center. (2001), 2001 Annual University Conference: "14 Rules For Writing Multiple-Choice Questions" [PDF document]. Available: https://testing.byu.edu/handbooks/14%20Rules%20for%20Writing%20Multiple-Choice%20Questions.pdf

[23] E. Erdfelder, F. Faul, and A. Buchner, "GPOWER: A general power analysis program," *Behavior Research Methods, Instruments, and Computers*, vol. 28, no. 1, pp. 1–11, Mar. 1996. https://doi.org/10.3758/bf03203630

Appendix A: Learning Objectives CSE2510

Course-level learning objectives:
After succesfully completing this course, the student is able to:
explain the basic concepts and algorithms of machine learning and their underlying statistical concepts.
implement, apply and evaluate basic ML algorithms in Python.
explain the concept of and identify (implicit) bias in data and ML algorithms.
Module-level learning objectives:
After practicing with the concepts of this week you are able to:
explain the basic ideas of machine learning and why and when it can be used,
explain the machine learning pipeline from data to training to testing and evaluation,
understand and apply the basic ideas of probability theory, decision theory, and Bayes' rule and their application in machine learning.
After practicing the topics taught this week, you should be able to:
1. explain how you obtain a classifier using a Gaussian (multivariate) distribution for each class
2. implement a simple univariate classifier in Python
3. explain what the 'curse of dimensionality' is
4. explain the advantages and disadvantages are of the Quadratic classifier, the LDA and the nearest mean classifier
5. identify when scaling of the features is important and how to cope with feature scaling
After practicing with the concepts of this week (part 1) you should be able to:
Explain the difference between parametric and non-parametric density estimation
Explain Parzen density estimation and classification
Explain k-nn density estimation and classification
Explain the advantages and disadvantages of Parzen en k-nn
Implement k-nn classifier in Python
Explain the Naive Bayes classifier, including the following:
– components and their function, independence assumption, dealing with missing data
– Continuous example, Discrete example
– Pros and cons
After practicing the topics taught this week (part 2), you should be able to:
Evaluate a trained Machine Learning model,
explain why a training and test dataset is needed,
explain what crossvalidation is, and bootstrapping
explain and compute learning curves
avoid overfitting by reducing the complexity of a classifier
explain confusion matrices and ROC curves
After this week you will be able to
Distinguish between generative and discriminative models
Reason about linear regression models and linear classifiers
Explain what hypothesis and cost functions are
Implement gradient descent to train a given linear model
Derive and implement logistic regression from its loss function
Identify the principles behind support vector classifiers
Describe some approaches to multi-class classification and their problems
After this lecture you are able to:
Explain the concept of (implicit) bias in data and algorithms of Machine Learning.
Argue about and pinpoint bias in a system/organisation taking into account technical, societal, legal, and/or educational points of view.
Apply the concept of (implicit) bias to machine learning applications.
After practicing with the concepts of this week you are able to
Explain when and why non-linear classifiers are needed
Explain the basic concepts of two non-linear classifiers: multi-layer perceptrons and decision trees
Explain the underlying algorithm of decision trees and (multi-layer) perceptrons and how they are trained.
Implement a decision tree
Explain why and how one can combine multiple classifiers
Contrast a decision tree and a random forest
After practicing the topics taught this week, you should be able to:
1. Differentiate between different types of dimensionality reduction techniques.
2. Motivate the choice of dimensionality reduction techniques over another.
3. Apply the PCA to reduce dimensionality of a given dataset
4. Differentiate between different clustering techniques
5. Motivate the choice of a clustering techniques over another.
6. Apply k-means clustering to a given data set
7. Apply hierarchical clustering to a given data set.

Figure 6: All learning objectives in CSE2510

The module-level learning objectives were topic-based, being either a weekly learning objective or per-lecture learning objective.

The first three highlighted learning objectives in Figure 6 were directly used in this study, while the last four highlighted learning objectives were used as inspiration for the creation of method two's learning objectives, which can be found in the last four rows of Figure 2.

Appendix B: Experiment, day one: elaborative interrogation group

Microsoft Forms of group 1: https://forms.office.com/e/g6XPHvM7Cx

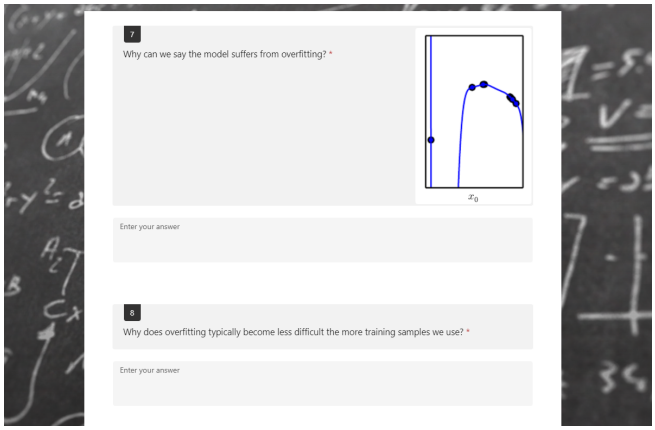

Figure 7: Windows Form used to conduct part 1 for the experiment group

Question	Answer
Q1	The five classes represent the five different species of penguin, the four features represent the measured characteristics of the penguin and the 800 objects represent the different penguins we're analyzing.
Q2	The arrows indicate the example's features, because they can be part of the measured characteristics of a plant.
Q3	The model suffers from underfitting as the function does not seem to follow the general trajectory of the points displayed in the graph.
Q4	The model suffers from overfitting as the function: first of all has a very complicated shape that is not needed to follow the points in the graph, Second of all hits all the points, making it seem very specific to this set.
Q5	Overfitting typically becomes less difficult the more training samples we use, as creating a function that follows all the training samples is quite complicated. It would require a lot of overtraining for our model to start doing this.
Q6	Classification is a supervised machine learning problem, as classification needs to train on labeled data to learn what classes to categorize new data as.
Q7	Our data could be overfitting. Despite its good performance on our training data, we have no information on how it performs on our test data. Thus, we have insufficient information to conclude anything.
Q8	We have previous labeled data of previous customers (as the company knows whether a customer has left or not) and we are trying to predict new customers. This is clearly a classification task.
Q9	Yes, recognizing the differences between written numbers can be done through k-NN by checking to see what the closest looking numbers of our current object look like.
Q10	Yes, recognizing the differences between written numbers can be done through Parzen by testing out different window functions and applying one accordingly.
Q11	Two answers can be counted as correct:
	k-NN, predicting whether a patient has an increased chance of heart disease can be done by looking at the characteristics of patients closest to ours and recommending based on that.
	Parzen, predicting whether a patient has an increased chance of heart disease can be done by looking at the characteristics of patients and prioritizing different qualities through different windows, we can predict an outcome based on that.
Q12	Two answers can be counted as correct:
	k-NN, predicting whether a user should be recommended a product can be done by looking at the characteristics of other users closest to ours and recommending a product based on the most similar purchasing history.
	Parzen, predicting whether a patient has an increased chance of heart disease can be done by looking at the different characteristics of the products and prioritizing different qualities through different windows, we can predict an outcome based on users with similar purchase history.

Figure 8: Answer sheet provided to both groups

Appendix C: Experiment, day one: control group

Microsoft Forms of group 2: https://forms.office.com/e/E5MK00hpsp

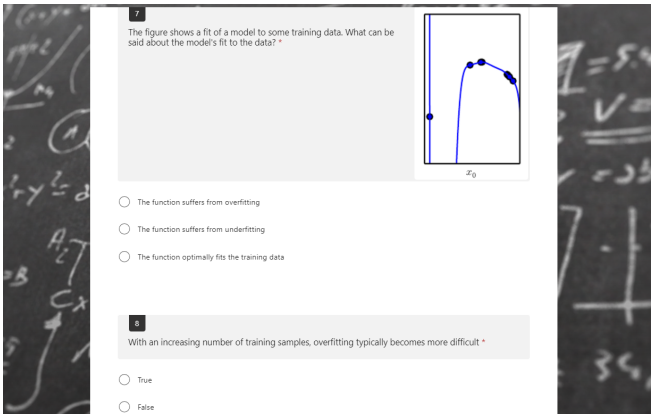

Figure 9: Windows Form used to conduct part 1 for the control group

Appendix D: Experiment, educational video

Educational video shown on day one of the experiment: https://www.youtube.com/watch?v=SRTmWvZ_iLo

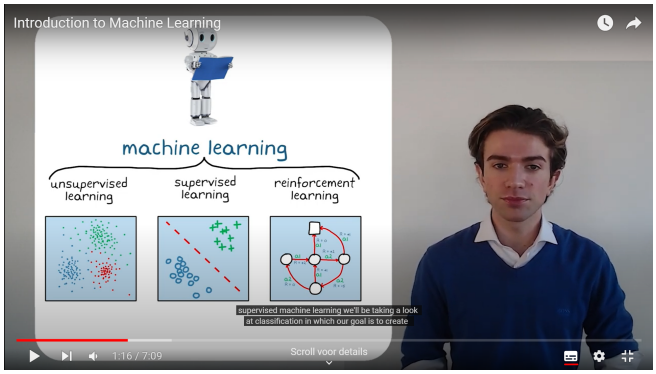

Figure 10: The educational video used in the experiment

Socrata educational video: https://www.youtube.com/watch?v=g7L_r6zw4-c&list=PLi01XoE8jYoi3SgnnGorR_XOW3IcK-TP6&index=2

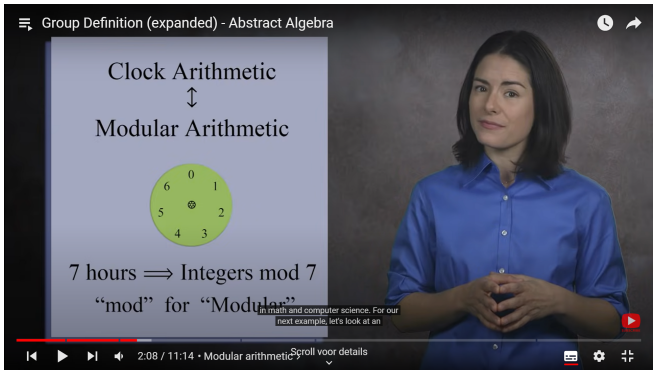

Figure 11: The educational video by Socrata used as inspiration [20]

TU Delft educational video: https://www.youtube.com/watch?v=envnkifm9lU

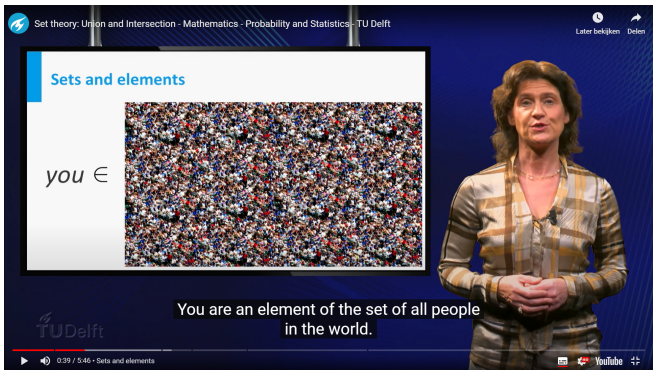

Figure 12: The educational video by TU Delft used as inspiration [19]

Appendix E: Experiment, day three

Microsoft Forms for day three: https://forms.office.com/e/bGUE0QHHTE

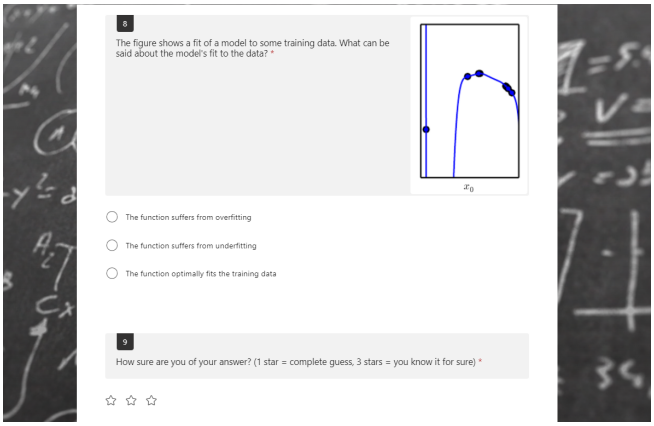

Figure 13: Windows Form used to conduct part 1

Appendix F: Power Analysis

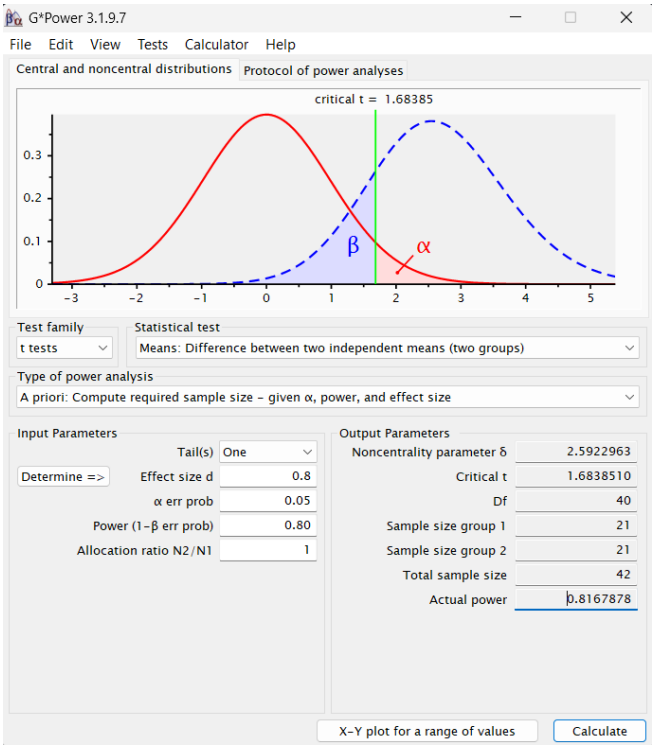

Figure 14: Sample size determination

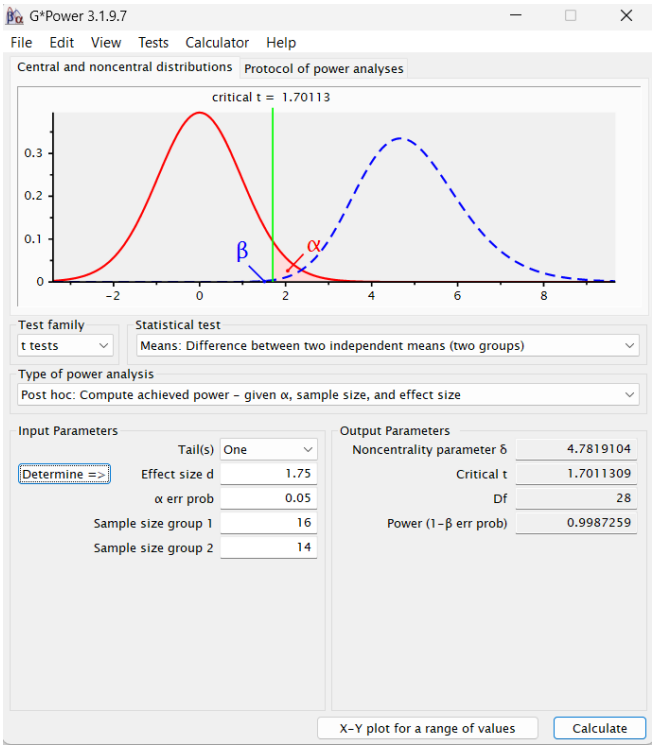

Figure 15: Achieved power

```
In [49]:
import scipy.stats as stats
group1 = [0.88, 0.94, 1.00, 0.97, 0.83, 0.73, 0.67, 0.56, 0.63, 0.85, 0.88, 0.87, 0.88, 0.67, 0.87, 0.92]
group2 = [0.67, 0.50, 0.70, 0.50, 0.56, 0.89, 0.59, 0.70, 0.59, 0.63, 0.67, 0.27, 0.44, 0.27]

t_statistic, p_value = stats.ttest_ind(group1, group2)

print(np.mean(group1), np.std(group1))
print(np.mean(group2), np.std(group2))

print("T-statistic:", t_statistic)
print("P-value:", p_value)

alpha = 0.05
if p_value < alpha:
    print("Reject the null hypothesis. There is a significant difference.")
else:
    print("Fail to reject the null hypothesis. There is no significant difference.")
mean_difference = np.mean(group1) - np.mean(group2)
pooled_std_dev = np.sqrt(((len(group1) - 1) * np.var(group1) + (len(group2) - 1) *
                          np.var(group2)) / (len(group1) + len(group2) - 2))
cohen_d = mean_difference / pooled_std_dev

print("Cohen's d:", cohen_d)

0.8218749999999999 0.1257089271889629 8
0.5700000000000001 0.16199647262473693 14
T-statistic: 4.624302785059956
P-value: 7.742570006193385e-05
Reject the null hypothesis. There is a significant difference.
Cohen's d: 1.7527732115791124
```

Figure 16: Mean, std, T-statistic, P-value and Cohen's d calculation