Enhancing Understanding in Receiver Operating Characteristic (ROC) Curve Analysis

An Investigation into the Impact of Interactive Teaching Methods

Alexandru-Sebastian Nechita[1]

Supervisor(s): Gosia Migut[1]

[1]**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
June 23, 2024

Name of the student: Alexandru-Sebastian Nechita
Final project course: CSE3000 Research Project
Thesis committee: Gosia Migut, Mark Neerincx

An electronic version of this thesis is available at http://repository.tudelft.nl/.

Abstract

The increasing demand for machine learning expertise calls for effective teaching methods for university-level courses. This research compares static versus interactive teaching methods in the context of machine learning, with the latter focusing on the student engaging more with the material. Specifically, this study investigates the impact of interactive visualisations on students' understanding of receiver operating characteristics (ROC) curve analysis, a critical concept in evaluating machine learning algorithms. Traditional static teaching methods often fall short of conveying complex ideas like ROC curves, which are pivotal in various fields, including medicine and psychology. This research also compares the efficacy of interactive versus static visualisations in enhancing student motivation. Twenty first-year computer science students from Delft University of Technology participated in the experiment and were randomly assigned to control (static visualisation) and experimental (interactive visualisation) groups. The results of the experiment were determined by analyzing the pre- and post-test scores, along with surveys measuring motivation. These indicate significant improvements in understanding for both groups, with a greater gain observed in the experimental group. This suggests that interactive visualisations may offer a superior approach to teaching complex machine learning concepts, but the experiment conducted in the study does not show a statistically significant difference between the static and interactive visualisations. The research also compares the student's motivation after completing an instructional material focused on the ROC, but the interactive visualisations did not provide better results. The study underscores the potential of interactive teaching tools to enhance educational outcomes in machine learning and highlights the need for further research into interactive methods for the teaching of machine learning.

1 Introduction

Machine learning is a rapidly growing field in both the tech industry and academia. As automation and data analysis become essential across various sectors, the demand for expertise in machine learning continues to increase. According to the Stanford AI Index 2024 report [1], AI-related degree programs are on the rise, and so is the need to efficiently and effectively teach machine learning courses. This increases the need to help students understand machine learning concepts.

This study compares two methods used in teaching machine learning: static and interactive. The first approach is characterized by teacher-centered methods such as lectures or slides and is commonly used in undergraduate programs. The second method involves interactive visualisations and may yield better results as effective visualisation reduces cognitive load by organizing information in a way that is easier for the brain to process [2]. According to Freeman et al. [3], active learning, which includes interactive elements, improves student performance across STEM disciplines, including computer science and engineering. This raises the question of whether these elements, such as interactive visualisations, specifically enhance understanding of machine learning concepts.

An important concept that students need to understand is how to evaluate a machine learning algorithm based on its performance. One method is the receiver operating characteristics (ROC) curve, which is a graph of the true positive rate instances against the false positive rate instances at each decision threshold setting. It was shown that the use of the ROC curve is useful in comparing and evaluating algorithms [4]. Not only that, but recent years have seen an increase in the use of ROC graphs in the machine learning community, as simple classification accuracy is often a poor metric for measuring performance [5].

The traditional teaching methods for ROC only contain static visualisations. This limits the student's ability to understand how the operating point is moving on the curve based on the decision threshold. In contrast, the interactive visualisation would provide the flexibility to adjust the threshold. Students would be able set various thresholds and compare how the true and false positive rates are changing. This facilitates the understanding of different scenarios such as classifying all the instances as a single class. Additionally, an effective interactive visualisation would allow the student to create imbalances in the class distributions and display the ROC curve in real-time.

While the ROC is an essential concept to learn, it is often overlooked as it is a challenging concept for students to understand [6]. Additionally, there is a significant gap in the studies addressing the teaching of ROC curves in university-level machine learning courses. This gap drives the need for research on teaching ROC in a machine-learning context. Hence, the research aim of this study involves trying to improve the teaching of ROC by using interactive visualisations [6].

1.1 Research Question

The main research question that grounds this investigation is as follows:

How do interactive receiver operating characteristics (ROC) curve visualisations compare with traditional static visualisations in terms of students' understanding of ROC analysis concepts?

In addition to this primary research question, the study explores:

- *Do students demonstrate an improved ability to interpret and apply ROC analysis concepts when using interactive ROC curve visualisations as opposed to static visualisations?*
- *Do students acquire a deeper understanding of ROC analysis concepts when taught using interactive visual-*

isations compared to traditional ones in terms of comprehension?

- *How do interactive teaching methods impact students' motivation after learning about the ROC curve?*

1.2 Research Paper Structure

To fully answer the above questions, the paper is structured into several sections. Section 2 discusses the background surrounding the research study. Next, the methodology of the study is fully detailed in section 3. Afterwards, section 4 gives a comprehensive description of the experimental setup. Then, section 5 will delve into the results for this study. Section 6 will discuss the responsible research done in the study, whereas section 7 will be reserved for discussions surrounding the study. Lastly, section 8 will conclude the paper.

2 Background

This section describes the background of the research study. The first subsection presents how a literature review was approached in order to answer the research questions. The next one explains the receiver operating characteristics curve along with the critical concepts needed to understand ROC. Lastly, the teaching of the ROC curve in the university setting is analysed, and the use of interactive visualisation is discussed.

2.1 Criteria for literature review

A literature review was performed using electronic databases such as Scopus, Google Scholar and IEEE Xplore, to find conference and research papers or academic books. The search contained the following keywords: "Receiver Operating Characteristics curve","ROC curve", "Area under the curve", "AUC", "interactive visualisations", "static visualisation","RIMMS" (Reduced Instructional Material Motivation Survey), "IMMS" (Instructional Material Motivation Survey), "motivation", "teaching" and "learning". Additionally, the references of identified articles were examined to further extend the literature coverage.

The selection process was based on several criteria: relevance to the research questions and emphasis on ROC teaching. The chosen sources include empirical studies that offer insights into the learning difficulties associated with ROC and case studies that demonstrate the effectiveness of visualisation techniques for improving comprehension.

2.2 Understanding of ROC analysis

ROC analysis is a well-explored topic in the literature, with a significant portion focusing on medical decision-making rather than machine learning. For example, Hanley and McNeil [7] and Metz [8] offer comprehensive explanations of ROC analysis, providing examples from the medical field. However, their discussions extend beyond the scope of an undergraduate machine learning course on ROC curves.

Despite this, the literature also includes research papers that acknowledge the utility of ROC in machine learning. Fawcett [5], for instance, introduces ROC graphs within the context of machine learning. He first covers the four outcomes of a binary classifier (true positive, false positive, true negative, false negative), with the first term indicating if the classification is correct and the second indicating the classification itself. Then, the true positive rate and false positive rate are explained. The exact formulas for these parameters are:

$$\text{True Positive rate} = \frac{\text{TP}}{\text{TP + FN}}$$

$$\text{False Positive rate} = \frac{\text{FP}}{\text{FP + TN}}$$

The ROC curve is a graph showing the performance of a classifier by displaying the true positive rate vs. the false positive rate at all classification thresholds. This graph can be used to calculate the area under the curve (AUC), which gives a summary of how well a model works across a variety of thresholds. Since the true and false positive rates are between 0 and 1, the AUC will also range in value from 0 (all predictions are incorrect) to 1 (all predictions are correct). Lastly, ROC graphs and AUC are useful tools for organizing classifiers and visualizing their performance.

2.3 Teaching of ROC curve

There is a notable gap in extensive research on teaching the ROC curve for machine learning at the university level. Nonetheless, Eng [6] describes how to create a learning experience that fosters an understanding of ROC curves through interactive laboratory exercises. Powell et al. [9] offer an innovative approach for educators to present the ROC curve to undergraduate students.

Little research concerning the types of visualisations used in ROC teaching was done. Yet, according to Naps and et al. [10], interactive visualisations appear promising in enhancing understanding computer science concepts, particularly when learners actively engage with them. For ROC, interactive visualisations would allow students to manipulate the data and parameters dynamically, providing immediate feedback on how changes impact the curve. Additionally, interactive tools can illustrate the effect of different classification thresholds in real-time, making it easier for students to grasp the concept of classifier performance across various scenarios. This level of engagement and immediate feedback is not possible with static visualisations.

3 Methodology

This section describes the methodology employed in the research study to fully understand the impact of interactive visualisation on student's level of comprehension and motivation after completing an instructional material on ROC curve. First, an experiment was designed, and participant selection criteria were established. Then, two learning objectives were defined, and instructional materials were developed accordingly. Finally, a post-experiment survey was conducted to assess the relationship between the type of visualisation used and students' understanding of the ROC curve.

3.1 Selection Criteria for Experiment

For this study, only undergraduate students were considered. Specifically, the participants are first-year computer science students enrolled at Delft University of Technology who did not complete a machine learning related course where ROC was taught. In addition, the students were asked in the survey if they completed a course of machine learning and if they had knowledge of ROC, and they were pre-tested on their knowledge. By doing so, the comparison between static and interactive visualisations can be correctly assessed.

Before taking part in the experiment, the students received an informed consent document detailing the research's objectives and potential risks. They were informed of their right to withdraw from the experiment at any time without any consequences. Finally, the participants were fully briefed on their involvement in the study [11].

3.2 Experimental Process

This subsection explores the steps taken to establish the good coordination for the experiment. Specifically, the first part describes the learning objectives desired for the experiment. The second involves the creation of Jupyter Notebooks that facilitate a deeper understanding of the topic through carefully structured content and visualisations.

3.2.1 Learning Objective Establishment

Before conducting the experiment, two digital notebooks explaining ROC curve analysis were prepared. Both notebooks contain similar material on ROC, differing only in the type of visualisation used. Additionally, the following two learning objectives were established for the notebooks:

1. Define ROC curves and recall their purpose in evaluating classification algorithms. (Knowledge)

2. Interpret ROC curves by understanding how changes in threshold values affect sensitivity and specificity. (Comprehension)

The learning objectives were realised using Bloom's taxonomy [12] which is a powerful tool for enhancing educational practices by promoting comprehensive learning. The objectives represent the first two stages of the taxonomy, specifically the knowledge and comprehension levels.

3.2.2 Notebook Creation

The notebooks were hosted on a Jupyter Notebook, with the material divided into multiple cells. Their creation adhered to the guidelines provided by Project Jupyter [13], specifically breaking down the content into smaller, manageable steps. For instance, the notebooks included a few exercises aimed at reinforcing the concepts presented.

Before introducing the notion of ROC curve, the notebooks introduce other essential concepts that are necessary for the understanding of the ROC curve. Firstly, the notebook starts by describing what a binary classifier is and provides a small example. Then, the classification outcomes such as true positive, false positive, true negative and false negative, are explained. Along with this, the notion of confusion matrix is also presented. Then, the evaluation metrics such as sensitivity, specificity and accuracy are explained. Then, the concept of ROC is introduced along with a comprehensive description of the ROC space.

The notebooks contain mainly text explaining the material, but it also consists of small instances of code. In order to discover the solution to the exercises provided, the students need to run certain cells. The code used in the notebooks is written in Python and the interactive visualisations were created with Matplotlib and Bokeh library.

The first plotting library was used for the first visualisation found in Figure 6 in Appendix A. Before seeing the visualisation, the student was presented with plot of positive and negative Covid instances based on temperature and oxygen level. The visualisation lets the student interact with the position of the decision boundary in the plot and visualise how the operating point is changing in the ROC curve.

The latter plotting library was chosen as it provides elegant and concise construction of versatile graphics and affords high-performance interactivity over large or streaming datasets. It was used for the next interactive visualisation, as depicted in Figure 1. In this case, the student could build their own class distribution for positive and negative instances by setting the mean and standard deviation of the distributions or the number of instances generated. Additionally, the student could interact with the decision threshold and see in real time that the operating point is moving on the curve. Lastly, the student can choose what plots to display, ranging from ROC curve, and area under the curve (AUC) plot to accuracy plot. Appendix A presents this visualisation in more detail in Figure 7 and Figure 8.

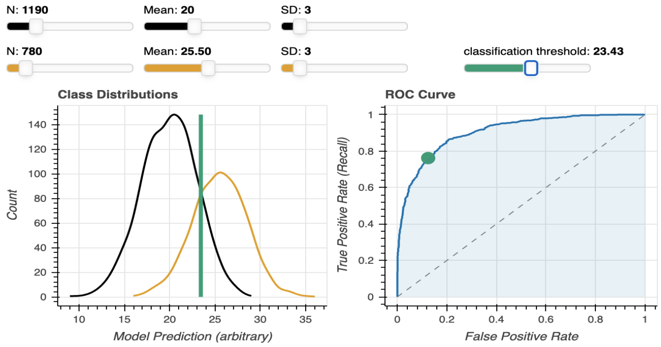

N: **1190** Mean: **20** SD: **3**
N: **780** Mean: **25.50** SD: **3** classification threshold: **23.43**

Figure 1: Interactive visualisation available in the notebook. It displays the class distributions for positive (orange) and negative (black) instances. N represents the number of data points generated by the normal distributions. The classification thresholds (green line) classifies the instance into positive and negative. The left side of the threshold is classified as negative, whereas the right side as positive. Mean is the mean of the distribution and SD is the standard deviation of the distribution.

The notebooks are publicly available for both interactive [1] and static visualisations [2], so any researcher could inspect them. In this way, the research is reproducible, and sets an important guide for anyone that is willing to continue the research on this topic. Lastly, the code used for the second visualisation, as seen in Figure 1 is partially taken and modified from another open-source project [14].

[1] https://github.com/SebiNechita/ROC_interactive

[2] https://github.com/SebiNechita/ROC_static

3.3 Survey Procedure

The objective of the survey was to identify any potential differences in understanding of the ROC curve. The first part of the survey included questions related to the material, while the second part featured statements about the students' motivation to learn the material. The content of the first part aligned with the two learning objectives, presenting various questions about ROC curve concepts. The first three questions, designed at the lowest level of Bloom's taxonomy, assessed whether students understood how the formulas for performance metrics are derived. The remaining six questions tested the next level in the taxonomy and were mainly taken from university-level exams. Each question was worth one point, and participants received a score based on their responses.

The survey was designed according to the Reduced Instructional Material Motivation Survey (RIMMS) [15] model to measure students' motivation after completing the digital notebooks. It included twelve statements focusing on the attention, relevance, confidence, and satisfaction of the students with the material. These four constructs represent the ARCS model [16] which constitutes a method for improving the motivational appeal of instructional materials. The statements utilized a 5-point Likert scale to assess student motivation. The student's responses are used to do quantitative analysis by calculating the mean scores for all four constructs.

4 Experimental Setup

This section details the experimental setup used for comparing the static and interactive visualisations. It includes the experiment design, sample size determination and experiment procedures.

4.1 Experiment Design

The research study uses a control-experimental group comparison design investigating the impact of interactive visualisation on the learning and comprehension of the concepts behind the ROC curve analysis. The design involves identifying independent and dependent variables. The independent variable is the use of interactive visualisation tools versus traditional learning methods. The dependent variables are learning outcomes and comprehension levels, assessed through the post-experiment survey and the students' motivation after completing the material.

In accordance with Campbell and Stanley's findings [17], participants were randomly assigned to either the control or experimental group. Additionally, pre-testing was employed for both groups to assess their initial knowledge and expertise, thereby minimizing potential biases. The questions used for the pre-testing and post-testing can be found in Appendix B.1.

4.2 Sample Size Determination

Before conducting the experiment, the sample size n of the experiment was calculated using the Cohen formula as it helps to determine the required sample size to detect an effect of a given size with a certain degree of confidence. The formula is:

$$n = 2\frac{(Z_{\alpha/2} + Z_{\beta})^2\sigma^2}{d^2}$$

In the formula mentioned above, the parameters were set as follows: the significance level $\alpha = 0.5$; the probability of Type II error $\beta = 0.2$; Z values correspond to critical values from the Standard Normal Distribution $Z_{\alpha/2} = 1.96$ and $Z_{\beta} = 0.84$; the effect size $d = 0.5$. These parameters would result in a sample size of approximately 64 participants. However, due to time constraints, it was decided that only twenty students will participate in the experiment. The number of participants was also influenced by the limitation of having only first-year students who are enrolled in the computer science undergraduate program. Future research should strive for a larger number of students to increase the validity of the results. The participants were randomly assigned to one of the notebooks, such that half of them were part of the control group and the other half part of the experimental group.

4.3 Experiment Setup

Each student was scheduled to participate individually in the experiment in a quiet environment. The first part of the experiment consisted of taking a pre-test of the knowledge. Afterwards, the participant had access to the notebook. There was not a specific time limit for the participants to read and understand the material provided to them. After the participants explored the content of the notebook, they were asked to partake in the post-experiment survey. Both the control and experiment group have access to the same content. However, the experimental group used the notebook with the interactive visualisations and the control group the one with the static visualisations.

5 Results

The findings that were discovered after the students finished the survey are examined in this part. The degree of knowledge the students have gained regarding the ROC curve is covered in the first subsection. The following one talks about the impact of the notebook visualisations on students' motivation.

5.1 Comprehension of the ROC curve

The experimental results from the interactive and static visualisations were analysed to detect if there exists a difference between the effect the two types of visualisation have on the comprehension of ROC curve concepts. The analysis of pre-test and post-test scores reveals a significant improvement in participant understanding across both visualisations.

The pre-test numbers are low for both the static and interactive visualisations because the participants were first-year undergraduate students who were not familiar with the ROC curve. In this way, it is tested if the participants did not have any previous knowledge of the subject. However, the post-tests demonstrate a greater understanding of the ROC curve concepts. This increase is statistically validated by a low p-values for both groups (e.g., the p-values are below 0.001). Figure 2 shows the results obtained by the control group which had access to the static visualisations, whereas figure 3 shows the results obtained by the experiment group which had access to interactive ones. The mean for the experimental group is 6.4 points from 9 possible points and the standard deviation is 1.77. The mean for the control group is 5.7 points and the standard deviation is 1.49.

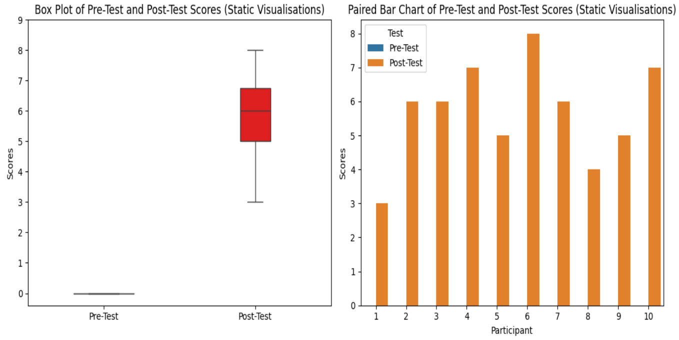


Figure 2: Pre- and Post-test scores for Static Visualisations. The left plot shows the means for the pre- and post-test scores. The right plot shows what scores each participant received in the post-test for the static visualisations

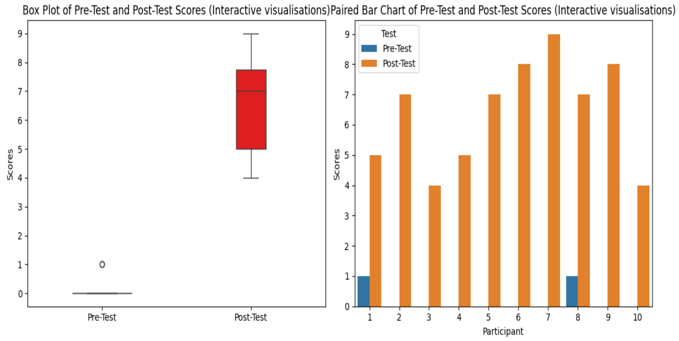


Figure 3: Pre- and Post-test scores for Interactive Visualisations. The left plot shows the means for the pre- and post-test scores. The right plot shows what scores each participant received in the post-test for the interactive visualisations

In order to correctly compare the knowledge gain for the two groups, the normalised gain was calculated for both groups using this formula:

$$\text{Normalised Gain} = \frac{\text{Post-test score} - \text{Pre-test score}}{\text{Max Score} - \text{Pre-test Score}}$$

For a better understanding of the results, the research uses statistical test such as t-test on the normalised gain for both groups. Firstly, the Shapiro-Wilk test is employed to verify the normality of the data. The comparison between the normalised scores of the groups could be visualised in figure 4. Then, t-test was used to verify if there is a statistically significant difference in the post-test scores. The obtained p-value from the t-test is 0.403. However, for a result to be considered statistically significant, the p-value should be below 0.05. This means that there is not statistically significant difference in the understanding of ROC curve analysis between the two groups.

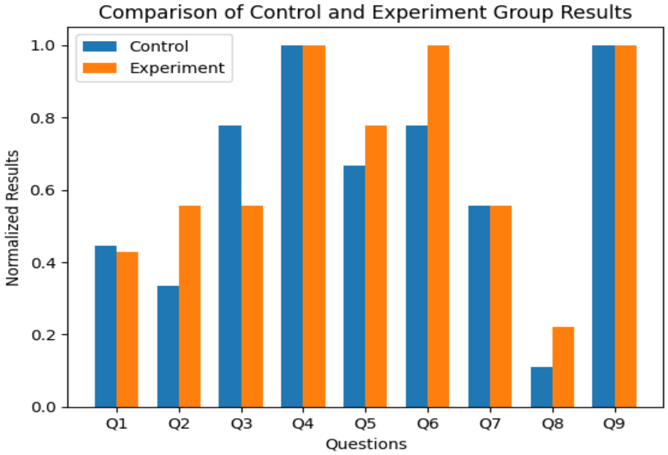


Figure 4: Comparison for the normalised score obtained for each question between control (static visualisations) and experimental (interactive visualisations) group

5.2 Student's motivation after completing the notebook

The ARCS model is used to evaluate and enhance motivational design in instructional materials. Using the Reduced Instructional Materials Motivation Survey (RIMMS), data was gathered on 12 statements categorized into the four dimensions of the ARCS model. For each statement, which can be found in Appendix B, the participants were asked to rate it with the help of 5-Likert scale. For every of the five options, some points were attributed. (e.g., "Not true" = 1 points, "Slightly true" = 2 point, "Moderately true" = 3 points, "Mostly true" = 4 points, "Very true" = 5 points). After the points were attributed, the attention, retention, confidence, satisfaction and the overall motivation was calculated for both groups. Figure 5 displays the results by the use of box plots.

In this analysis, the Mann-Whitney U test was employed for all comparisons due to the non-normal distribution of at least one group in each dataset. The results indicate a statistically significant difference in overall motivation between the interactive and static groups (Statistic=6166.5, p-value=0.0397). However, when examining the individual components of motivation, no significant differences were found: attention (Statistic=420.0, p-value=0.6367), retention (Statistic=367.5, p-value=0.1897), confidence (Statistic=392.5, p-value=0.3652), and satisfaction (Statistic=363.0, p-value=0.1738) all showed p-values greater than 0.05. These results suggest that while the aggregated motivation scores differ significantly between the groups, the individual constructs of attention, retention, confidence, and satisfaction do not exhibit significant differences on their own. This discrepancy highlights the potential impact of aggregation in detecting overall motivational differences that may not be apparent when examining the constructs separately.

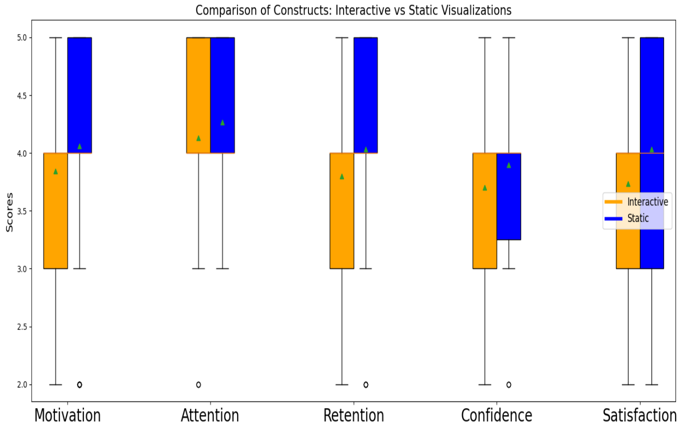


Figure 5: Box and Whiskers plots comparing static and interactive visualisations in terms of student's motivation. The first aspect is measured is the motivation as a whole, whereas the other four represent the four specific constructs needed to measure motivation: attention, retention, confidence and satisfaction

6 Responsible Research

This section illustrates the practices employed during the study to maintain and safeguard the integrity of the research. The first part of the section concerns the data collection and methods that ensure the safety of the private data. The second part reflects on the reproducibility of the study.

6.1 Data collection

The research study adheres to the Netherlands Code of Conduct for Research Integrity, ensuring the protection and privacy of participants' data through multiple measures. Each participant was informed about the study's goals and risks through an Informed Consent document. The experiment has undergone a review by the Human Research Ethics Committee, which included a risk assessment detailing potential hazards and their mitigation strategies, as well as a data management plan describing the secure processing and storage of data to ensure reproducibility.

6.2 Reproducibility

The research was designed with reproducibility as a key focus throughout its duration. Both the experiment and the surveys are fully replicable, with detailed explanations provided in the methodology section. The code for the notebooks is publicly available and anyone can access and interact with the visualisation. Thorough descriptions of the survey methods, experimental setup, and data collection procedures were included to enable other researchers to replicate the study. These details cover the specific survey questions, the data collection timeline, and the statistical methods used in data analysis. Appendix B contains the questions and statements used in the survey.

7 Discussion

The results indicated that both interactive and static visualisations improved students' understanding of ROC curve concepts. However, the observed mean score for the experimental group, which used interactive visualisations, was higher than that of the control group. This suggests that interactive visualisations may offer additional benefits over static visualisations, even though the difference was not statistically significant.

The lack of statistical significance might be attributed to the small sample size, which reduces the power to detect subtle differences. The small number of participants could be attributed to short duration of the research and the decision of having only first-year students for Delft University of Technology. While this decision ensured that the participants would have the necessary knowledge to handle a Jupyter Notebook and understand the necessary mathematics for ROC curve, it greatly limited the size of the groups. Nonetheless, the higher mean score and positive feedback from students in the experimental group suggest that interactive visualisations could enhance comprehension, consistent with the principles of active learning.

The results from the motivational aspect of the study yielded a surprising outcome: static visualisations outperformed interactive ones in terms of student motivation. Across all four constructs measured, the control groups scored higher than the experimental group. However, both groups showed a high level of motivation, suggesting that the type of visualisation may not significantly impact student motivation. It is important to note that the students were tested on only two learning objectives, corresponding to the two lower levels of Bloom's taxonomy, and more complex tasks at higher taxonomy levels were excluded. This exclusion aimed to make the notebooks quicker to complete and to reduce cognitive load, resulting in the absence of exercises involving ROC curve coding. Further research is necessary to understand the impact of interactive visualisations on student motivation when engaging with instructional materials that include more complex learning objectives.

8 Conclusions and Future Work

The research study is addressing the effectiveness of interactive and static visualisation on the understanding of the receiver operating characteristics (ROC) curve, as well on the student motivation after completing instructional material with the visualisations. The study involves a comparative experiment between the two types of visualisation.

The results indicated a general improvement in understanding ROC curves across both groups—those exposed to interactive visualisations and those exposed to static visualisations. Both groups demonstrated a significant increase in post-test scores compared to their pre-test scores, indicating that the teaching materials were effective overall. However, the difference between the two groups was not statistically significant, as evidenced by the p-value of 0.403, which is above the 0.05 threshold for statistical significance. Additionally, the research also compares the student's motivation after completing the notebooks, but the interactive visualisations did not provide better results.

Despite the lack of a statistically significant difference between the groups, the higher mean score observed in the experimental group suggests a potential benefit of interactive visualisations. These findings are consistent with previous research indicating the positive impact of interactive learning methods on student performance in STEM disciplines. The interactive visualisations allowed students to manipulate decision thresholds and observe real-time changes in the ROC curve, potentially offering a more intuitive grasp of the concepts.

This study contributes to the limited research on teaching ROC curve analysis within undergraduate machine learning courses. By developing and evaluating interactive visualisation tools, this research provides a framework that other educators can replicate and build upon. The visualisations provides the opportunity of adjusting the decision thresholds, making it easy to test different scenarios. The notebooks created for this study, which are publicly available, offer a valuable resource for enhancing receiver operating characteristics curve education. Both notebooks are accessible to students enrolled in computer science undergraduate programs and could be used as resource in introductory machine learning courses.

The study's sample size was relatively small, with only twenty participants, which has impacted the ability to detect statistically significant differences. There were several factors that affected the sample size. Firstly, it has taken each participant about thirty minutes to finish their notebooks. First-year students were less inclined to participate in the experiment as a result. Secondly, the experiments were not scalable, so each student participated at a different time schedule. Lastly, the factor which represents the biggest limitation is that participants had to be first-year computer science students. However, the decision to solely consider first-year students ensured that the participants had the necessary mathematical expertise to understand the instructional content and that they did not complete the machine learning course offered in the second year of bachelor. Future research should consider larger sample sizes and diverse student populations to validate these findings further. Additionally, exploring interactive visualisation tools for other machine learning concepts could enhance the quality of the education in the field and this study shows this potential.

To overcome the limitations mentioned above, more learning objectives should be considered for guiding the material. For instance, the other levels of Bloom's taxonomy should be included to further inspect the differences between the types of visualisations used in the notebooks. This could include programming exercises or analysing various ROC curves. Future work could also investigate the impact of visualisations when ROC curves are used for specific binary classifiers such as support vector machines or neural networks, but this requires the participants to have a good understanding of machine learning principles. Lastly, future research could replicate the experiment for complex concepts such as the multiclass ROC curve.

In conclusion, this study emphasises the potential of interactive visualisations to enhance the learning experience in machine learning education. While the difference in comprehension between interactive and static visualisations was not statistically significant for the conducted experiment, the observed trends warrant further investigation. Interactive teaching methods, when carefully designed, hold promise for making complex concepts like ROC analysis more accessible and engaging for students. This demonstrates that interactivity is a useful tool for machine learning education.

References

[1] A. I. Team, "Ai index stanford report 2024," tech. rep., Stanford University, 2024. Available online.

[2] C. Ware, *Information Visualisation: perception for design*. S.L.: Morgan Kaufmann, 2020.

[3] S. Freeman, S. L. Eddy, M. McDonough, M. K. Smith, N. Okoroafor, H. Jordt, and M. P. Wenderoth, "Active learning increases student performance in science, engineering, and mathematics," *Proceedings of the National Academy of Sciences*, vol. 111, no. 23, pp. 8410–8415, 2014.

[4] K. A. Spackman, "Signal detection theory: Valuable tools for evaluating inductive learning," in *Proc. Sixth Internat. Workshop on Machine Learning*, (San Mateo, CA), pp. 160–163, Morgan Kaufman, 1989.

[5] T. Fawcett, "Introduction to roc analysis," *Pattern Recognition Letters*, vol. 27, pp. 861–874, 06 2006.

[6] J. Eng, "Teaching receiver operating characteristic analysis: an interactive laboratory exercise," *Academic Radiology*, vol. 19, no. 12, pp. 1452–1456, 2012.

[7] J. Hanley and B. McNeil, "The meaning and use of the area under a receiver operating characteristic (roc) curve," *Radiology*, vol. 143, no. 1, p. 29 – 36, 1982.

[8] C. E. Metz, "Basic principles of roc analysis," *Seminars in Nuclear Medicine*, vol. 8, no. 4, pp. 283–298, 1978.

[9] L. Powell, R. Mariani, G. Powell, C. A. Al, and Y. Chen, "Innovative approach for teaching the receiver operating characteristic (roc) curve," *Issues in Information Systems*, vol. 24, no. 3, p. 97 – 104, 2023. Cited by: 0; All Open Access, Gold Open Access.

[10] T. L. Naps, R. Fleischer, M. McNally, G. Rößling, C. Hundhausen, S. Rodger, V. Almstrum, A. Korhonen, J. Velázquez-Iturbide, W. Dann, and L. Malmi, "Exploring the role of visualization and engagement in computer science education," p. 131 – 152, 2002. Cited by: 314; All Open Access, Green Open Access.

[11] M. Mauthner, M. Birch, J. Jessop, and T. Miller, *Ethics in Qualitative research*. 1 2002.

[12] D. R. Krathwohl, "A revision of bloom's taxonomy: An overview," *Theory into Practice*, vol. 41, no. 4, p. 212 – 218, 2002. Cited by: 4011; All Open Access, Green Open Access.

[13] L. A. Barba, L. J. Barker, D. S. Blank, J. Brown, A. B. Downey, T. George, L. J. Heagy, K. T. Mandli, J. K. Moore, D. Lippert, K. E. Niemeyer, R. R. Watkins, R. H. West, E. Wickes, C. Willing, and M. Zingale, "Teaching and learning with jupyter." Online, December 6 2019. Available: https://jupyter4edu.github.io/jupyter-edu-book/.

[14] D. H. Brown, “Interactive classification metrics.” https://github.com/davhbrown/interactive-classification-metrics, 2024.

[15] N. Loorbach, O. Peters, J. Karreman, and M. Steehouder, “Validation of the instructional materials motivation survey (imms) in a self-directed instructional setting aimed at working with technology,” *British Journal of Educational Technology*, vol. 46, no. 1, pp. 204–218, 2015.

[16] J. M. Keller, “Development and use of the arcs model of instructional design,” *Journal of instructional development*, vol. 10, pp. 2–10, Sep 1987.

[17] D. T. Campbell and J. C. Stanley, *Experimental and quasi-experimental designs for research*. Belmont Wadsworth, 1963.

A The Jupyter Notebook

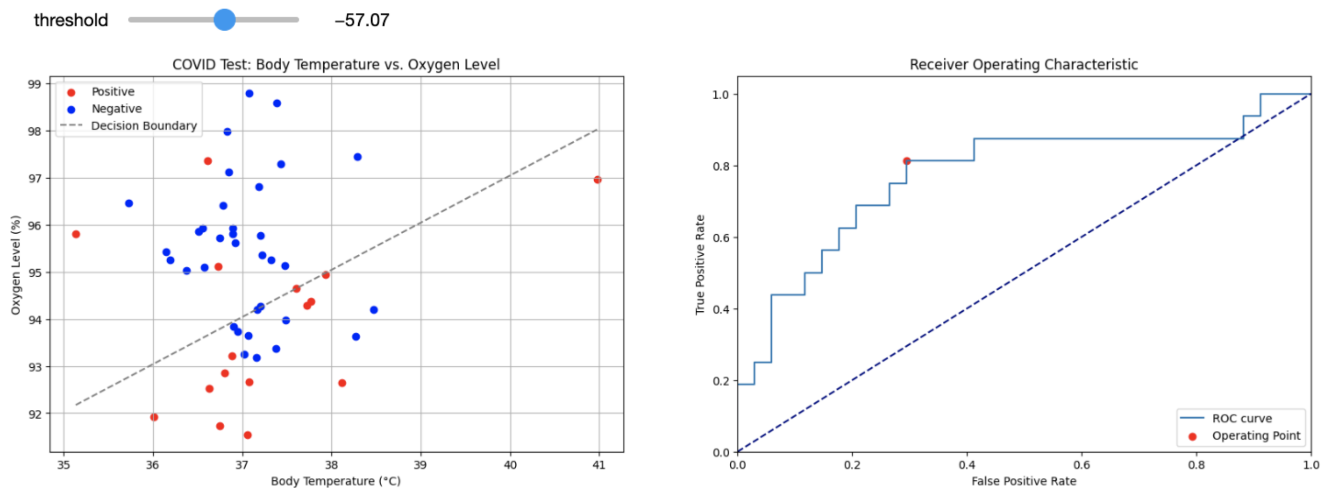
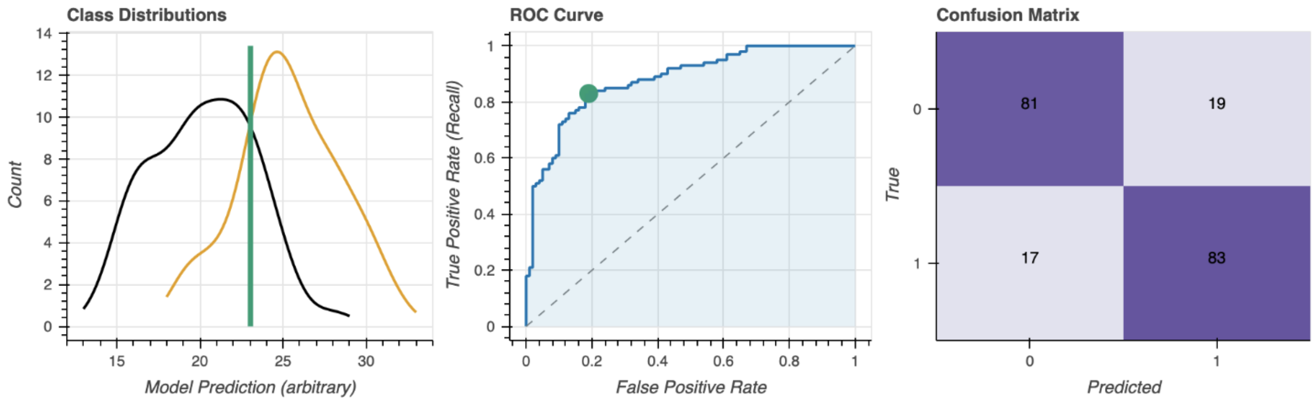


Figure 6: First interactive visualisation present in the notebook. The left plot shows data points, representing patients being tested for Covid, which are of two types: positive and negative. The y-axis represents the oxygen level of a patient and the the x-axis is the body temperature. The dotted line represents the decision boundary and is used to determine the operating point in the ROC graph displayed in the right plot.



	0	1
0	81	19
1	17	83

Figure 7: Second interactive visualisation present in the notebooks along with the corresponding confusing matrix. The leftmost plot depicts two class distributions, the black one being the negative instances whereas the orange one being the positive instances. The classification thresholds (green line) classifies the instance into positive and negative. The left side of the threshold is classified as negative, whereas the right side as positive. The next plot depicts the ROC graph for the corresponding class distributions. The rightmost shows the confusing matrix with the y-axis representing the actual values and x-axis being the predicted values.

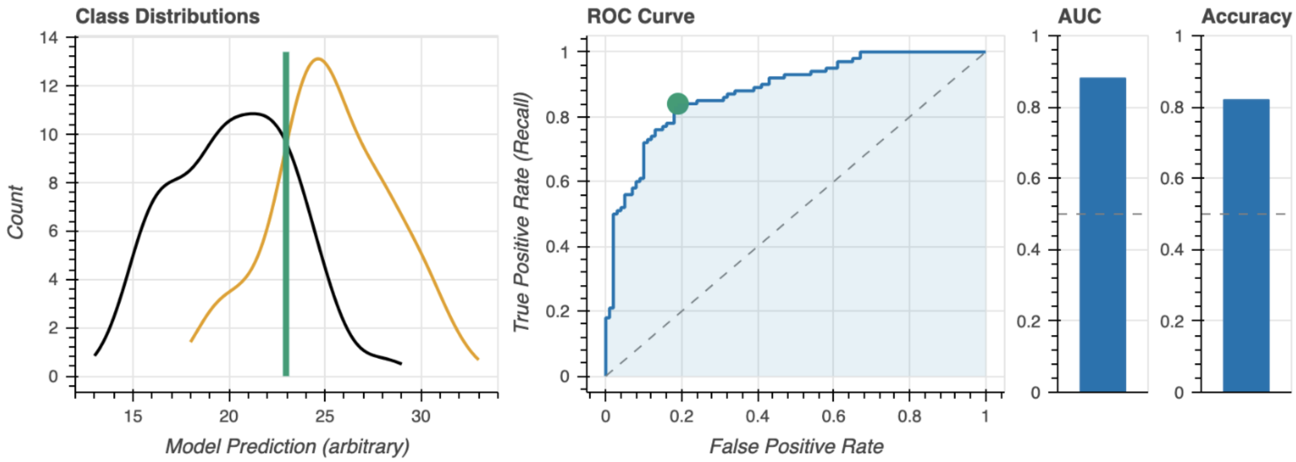


Figure 8: Second interactive visualisation present in the notebooks along with two plots representing the corresponding area under the curve and accuracy. The leftmost plot depicts two class distributions, the black one being the negative instances whereas the orange one being the positive instances. The classification thresholds (green line) classifies the instance into positive and negative. The left side of the threshold is classified as negative, whereas the right side as positive. The next plot depicts the ROC graph for the corresponding class distributions.

B Post-Experiment Survey

B.1 Theory Questions

1. Can you express the sensitivity of a classifier as an expression of TP, FP, TN, FN?
2. Can you express the specificity of a classifier as an expression of TP, FP, TN, FN?
3. Can you express the accuracy of a classifier as an expression of TP, FP, TN, FN?
4. What do TPR and FPR mean?
   - TPR is the number of correct responses. FPR is the number of incorrect responses.
   - TPR is the proportion of answers that were provided correctly as 'true'. FPR is the proportion of answers that were provided incorrectly as 'true'.
   - TPR is the proportion of answers that were provided correctly as 'true'. FPR is the proportion of answers that were provided incorrectly as 'false'.
   - I don't know!
5. What are on the x and y axes in an ROC plot?
   - x-axis: FP rate, y-axis: TP rate
   - x-axis: Number of FPs, y-axis: Number of TPs
   - x-axis: Number of TPs, y-axis: Number of FPs
   - I don't know!
6. What does area under the curve for an ROC plot tell us?
   - How well the model works at its optimum decision threshold
   - Which is the optimum decision threshold?
   - It gives a summary of how well a model works across a variety of thresholds.
   - I don't know!

7. The imagine illustrates an ROC curve plot, the AUC and the accuracy for a particular threshold. Which of the following statements are correct?
   - There are significantly more positive instances than negative instances and they are all classified as negative.
   - There are significantly more negative instances than positive instances and they are all classified as negative.
   - There are significantly more positive instances than negative instances and they are all classified as positive.
   - There are significantly more negative instances than positive instances and they are all classified as positive.
   - I don't know!
8. Given a decision function g(x) for a binary classifier, consider the example of an ROC curve as given in the figure. When we increase the threshold used for the classifier, what typically happens to the sensitivity (true positive rate) and specificity (true negative rate) of the classifier?
   - The sensitivity will increase, the specificity will decrease.
   - The sensitivity will decrease, the specificity will increase.
   - Both will increase.
   - Both will decrease.
   - I don't know!
9. What is the AUC?

B.2 Motivation Section

These are the statements used in the motivation side of the survey. At the end of each statement, it is mention the construct measured by the statement.

1. As I worked on this lesson, I was confident that I could learn the content. (Confidence)
2. It is clear to me how the content of this material is related to things I already know. (Retention)
3. The way the information is arranged on the pages helped keep my attention. (Attention)
4. I enjoyed this lesson so much that I would like to know more about this topic. (Satisfaction)
5. I really enjoyed studying this lesson. (Satisfaction)
6. The good organization of the content helped me be confident that I would learn this material. (Confidence)
7. After working on this lesson for a while, I was confident that I would be able to pass a test on it. (Confidence)
8. The content and style of writing in this lesson convey the impression that its content is worth knowing. (Retention)
9. The quality of the writing helped to hold my attention. (Attention)
10. The content of this lesson will be useful to me. (Retention)
11. The variety of reading passages, exercises, illustrations, etc., helped keep my attention on the lesson. (Attention)
12. It was a pleasure to work on such a well-designed lesson. (Satisfaction)