Measuring students' progress in Machine Learning

A case study of Decision Trees and Random Forests

Calin Manoli

Responsible Professor and Supervisor: Gosia Migut

[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 25, 2023

Name of the student: Calin Manoli
Final project course: CSE3000 Research Project
Thesis committee: Gosia Migut, Christoph Lofi

Abstract

Machine Learning (ML) is a rapidly growing field, therefore ensuring that students deeply understand such concepts is of key importance in order to certify that they are prepared for the challenges and opportunities of the future workforce. Despite this, literature on teaching ML and assessing students' understanding with regard to this field is scarce. Hence, this research aims to provide an extensive analysis of the best practice within the ML field, with the main focus of the study being the decision trees and random forests classifiers. An analysis of learning outcomes is conducted using Bloom's taxonomy, guidelines for creating assessments that reflect students' understanding levels are provided and a series of interviews and surveys are conducted in order to analyze the need for certain questions during the course examination. The results are then analyzed and key findings such as the need to structure the course such that decision trees are assessed as a prerequisite for learning random forests are further discussed. The research is concluded with a set of recommendations that could be integrated into future editions of the course in order to assess student progress in a more efficient manner.

1 Introduction

According to the IBM Global AI Adoption Index (2022) [7] Machine Learning (ML) is rapidly changing almost every industry and currently impacts numerous aspects of our lives. However, teaching students ML effectively is not a trivial task. ML algorithms are very complex and require students to understand multiple related subjects, such as Linear Algebra, Probability and Statistics, as well as being able to write high-quality code. In order for teachers to be able to effectively teach students ML, good pedagogical approaches, efficient assessments that measure student understanding for each particular sub-topic, and ways to create and provide constructive feedback are recommended [8] [18].

Naturally, teachers should be able to accurately assess student progress in order to uncover common misunderstandings amongst students as well as provide better feedback and more accurate grades. The focus will be on the decision trees and random forests classifiers module of the course. We chose this module due to the fact that decision trees are believed to be easier to understand by learners [5], they are efficient for large quantities of data [14] and have practical uses in fields such as Data Mining [15], which is also part of the Computer Science & Engineering curriculum at TU Delft, hence students would benefit of the possible module's structural improvements throughout their studies. Yet another reason for choosing this module is the fact that decision trees and random forests are closely related, with the latter extending on the complexity of the concepts presented by the former, hence making decision trees required for effectively learning random forests, which essentially encapsulates the background motivation for the research in a more accessible manner: How can teachers ensure that students present the prerequisite knowledge for being able to proceed to learn more complex concepts that extend upon the previously learned material?

Formally, this research will focus on answering the following question:

> *How can we programmatically review the decision trees and random forests module such that teachers can efficiently ensure that students thoroughly understand the most important concepts and are prepared to use the learned material in the future?*

Previous research has been conducted on this topic and showed potential designs of ML courses that facilitate teaching. Wangenheim et al. [6] showed potential ways to teach ML to students within primary and secondary school. Although this research is relevant and reveals important information, our aim is to study the progress in learning ML for university students, therefore we will use and extend upon this research. Sulmont et al. [17] uncovered how teachers can make use of the SOLO taxonomy in order to facilitate student learning. The research presents useful guidelines that can be used in order to use taxonomies to analyze, improve and create new learning objectives. Nonetheless, we advocate for the use of Bloom's taxonomy instead of the SOLO taxonomy, hence we will adapt this research to match our choice of taxonomy. Further justification for this choice is described more thoroughly in the ”Evaluating learning objectives” section.

Formally, this paper will focus on answering sub-questions such as:

1. What are the current formulated learning outcomes of the decision tree and random forests classifiers module and how can they be programmatically analyzed and extended in order to better show the teachers' expectations from students?
2. How can the formulated learning objectives be used to create and redesign exam questions and assessments in order to measure student progress more efficiently?
3. What are the opinions and insights of both teachers and teaching assistants and how can we use them in the assessment redesign process?

The main contributions of this article are to provide a replicable analysis of the decision trees and random forests module while extending upon the formulation of learning objectives and providing systematic means for the assessment creation process. Another key contribution is providing valuable survey insights, followed by discussions and possible proposals based on the results.

In order to achieve the previously defined goals of this research, the paper is structured as follows: The Methodology section provides more details regarding the chosen methods for this research and how these methods can be used to recreate it. The Evaluating learning objectives section discusses the reasoning behind choosing Bloom's taxonomy as the main analysis framework and how it can be used, while also proposing possible additions to the existing list of learning objectives previously defined by the ML course staff. This section aims at answering the first sub-question mentioned above. The Designing effective assessments section provides guidelines for a systematic way of creating assessments based on the learning objectives of the course. This section also discusses different possible assessment question formulations which are later analyzed by Teaching Assistants (TAs) with the help of surveys. Hence, by the end of this section, the second sub-question is answered. The Interviews and Surveys: Setup and Results section provides an overview of the questions professors and TAs were asked and analyzes their results. The Discussion section further analyzes the results and yields possible interpretations, while comparing them with related previous work. By the end of the last two aforementioned sections, the third and last sub-question is answered. The Responsible Research section reflects on the ethical aspects and concerns regarding this paper. The Future Work section proposes future research that is needed on this subject as well as discusses future possible improvements. Lastly, the Conclusion section summarizes the most important aspects and findings of this research.

2 Methodology

This section discusses the methods used for collecting and analyzing data. The aim of this section is to provide a systematic overview of the methods used in order to complete this research, such that the process is as repeatable as possible in future studies.

2.1 Learning objectives review

The first method used within this research was using Bloom's taxonomy as a framework for analyzing the learning objectives for the modules in discussion. This was the preferred framework for conducting this type of analysis since it proposes six levels of understanding which can be easily and consistently correlated with the learning outcomes formulation. Further justification for choosing this framework is described in the "Evaluating learning objectives" section. As such, the initial part of the process involved making use of the cognitive levels defined by the framework in order to analyze and extend the learning objectives already defined by the course staff. Therefore, after thoroughly analyzing the lectures and material presented during the course with regard to Bloom's taxonomy, we suggest a series of additions that we believe, based on related literature [16] [8], that would support the teachers in creating a better-defined overview of the knowledge that students should possess in order to be considered sufficient.

2.2 Designing alternative exam questions

The second phase of the research consists in creating new assignments and exam questions in order to support professors in recognizing student progress with regards to the decision trees module more efficiently. The creation process was based on previous research in this domain. Although research discussing assessment design specifically for ML courses is scarce, there is enough research conducted on teaching Computer Science as a whole [11], [16], [3], hence it was reviewed for the scope of this paper. In doing so, we present a mapping that can be used to programmatically correlate the most efficient exam question type, given the learning objective description.

2.3 Conducting interviews and surveys

The third and last phase of the research involves conducting and analyzing interviews and surveys with professors and Teaching Assistants (TAs) who have taken part in helping students understand the concepts taught in the ML course. The surveys were designed to assess the reliability and validity of possible alternative exam questions and assignments that we have previously created throughout this research. We decided that both the interview and survey questions must be included in their corresponding Appendix sections of this paper and the anonymized answered transcripts will only be made available upon request. This decision is further discussed in the "Responsible research" section.

3 Evaluating learning objectives

According to D. Kennedy [8], learning objectives are used to clearly define what the professors expect from students in order to consider their learning successful. Students can also benefit from having clearly defined learning objectives as they can understand what is going to be expected of them throughout the course as well as have a better overview of their overall preparedness. Naturally, learning objectives can have different levels of difficulty. For example, students might only be required to remember certain concepts of the course, while having to be able to implement others by themselves. Therefore, it is important that we understand and are able to analyze different learning objectives in order to be able to categorize them based on how "deeply" students need to understand certain topics. This analysis can later be used in order to create effective assessments that test whether students were able to meet professors' expectations, a process which is further discussed in the following section. A thorough analysis of the current learning outcomes is, therefore, recommended since, according to C. Starr et al. [16], creating quality learning objectives can help create effective assessments programmatically, while also improving communication amongst academics working on curriculum development.

Several frameworks have been proposed in order to effectively categorize learning objectives based on the levels of understanding that students need to reach in order to achieve the expected knowledge. However, most literature suggests that Bloom's taxonomy should be used in order to achieve this goal [12] [8] [11].

Bloom's taxonomy [1] is a collection of three hierarchical models used to categorize educational learning objectives into levels of complexity and specificity. The learning objectives for the cognitive, emotional, and psychomotor domains are covered by the three lists. The cognitive domain is typically used to organize curricular learning objectives, evaluations, and activities. Hence, this research will also make use of the cognitive component of Bloom's taxonomy in order to categorize and analyze the learning objectives presented by the ML course staff.

As explained in the Introduction section, only the decision tree classifier and the random forest classifier will be taken

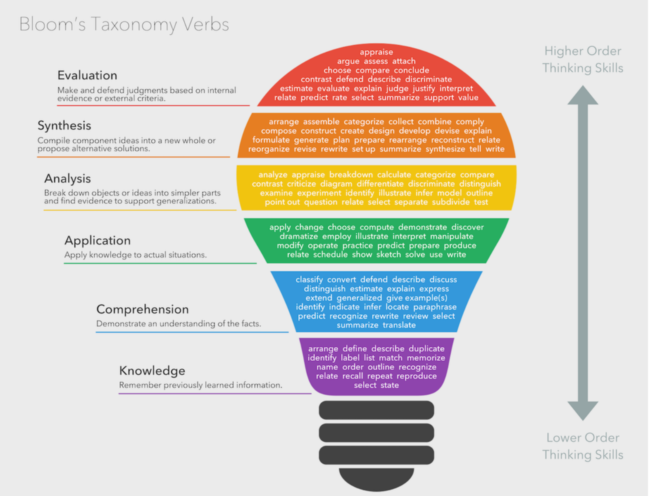

Bloom's Taxonomy Verbs
Evaluation
Make and defend judgments based on internal evidence or external criteria.
appraise argue assess attach choose compare conclude contrast defend describe discriminate estimate evaluate explain judge justify interpret relate predict rate select summarize support value
Synthesis
Compile component ideas into a new whole or propose alternative solutions.
arrange assemble categorize collect combine comply compose construct create design develop devise explain formulate generate plan prepare rearrange reconstruct relate reorganize revise rewrite set up summarize synthesize tell write
Analysis
Break down objects or ideas into simpler parts and find evidence to support generalizations.
analyze appraise breakdown calculate categorize compare contrast criticize diagram differentiate discriminate distinguish examine experiment identify illustrate infer model outline point out question relate select separate subdivide test
Application
Apply knowledge to actual situations.
apply change choose compute demonstrate discover dramatize employ illustrate interpret manipulate modify operate practice predict prepare produce relate schedule show sketch solve use write
Comprehension
Demonstrate an understanding of the facts.
classify convert defend describe discuss distinguish estimate explain express extend generalized give example(s) identify indicate infer locate paraphrase predict recognize rewrite review select summarize translate
Knowledge
Remember previously learned information.
arrange define describe duplicate identify label list match memorize name order outline recognize relate recall repeat reproduce select state
Higher Order Thinking Skills
Lower Order Thinking Skills

Figure 1: Visualisation of Bloom's taxonomy [4].

into consideration and analyzed from the non-linear classifiers module of the course.

The learning outcomes taken from the ML course material are the following:

1. **Explain** when and why non-linear classifiers are needed (Synthesis)
2. **Explain** the basic concepts of the decision trees classifier (Comprehension)
3. **Explain** the underlying algorithm of decision trees and how they are trained (Comprehension)
4. **Implement** a decision tree (Application)
5. **Explain** why and how one can combine multiple classifiers (Synthesis)
6. **Contrast** a decision tree and a random forest (Analysis)

Figure 1 includes the verbs used to describe learning outcomes corresponding to each category. The verbs used to match the learning outcome with its corresponding Bloom category have been written in bold font, while the category has been added after the learning objective within parentheses.

Although Bloom's taxonomy provides a relatively consistent framework for analyzing the learning outcomes, there are certain inherent limitations to it. For example, the verb "Implement" is not part of the figure, however, we decided it would belong to the Application category, since, in order to implement a classifier, students are essentially required to be able to apply previously gained knowledge within an actual program/situation, which precisely matches with the definition of the Application cognitive level. This is also consistent with K. Cakıcı's work regarding learning objective analysis within the ML course at TU Delft [20], as the same decision has been made throughout his research. Another limitation is the overlap of certain verbs with multiple cognitive categories. The word "explain" belongs to both the "Synthesis" category as well as the "Comprehension" category. It can be seen that both the first and the second learning objectives make use of the verb "explain", however, these objectives aim to test different levels of understanding. In order to correctly determine which cognitive category each objective corresponds to, we consulted with an experienced professional within the teaching staff of the ML course.

While using this framework in analyzing the learning objectives presented in the module material, we uncovered several factors that could be improved. For example, there is also only one learning objective that focuses on students' understanding of the random forest classifier. Moreover, the initial learning outcomes did not include any objectives regarding key elements that define how decision trees work, such as the information gain function. Hence, we propose the addition of the following learning objectives:

- Explain the basic concepts of the random forests classifier.
- Implement a random forest classifier.
- Describe the information gain function.
- Describe different node-splitting criteria (Misclassification, Entropy, Gini Index)
- Compute the confusion matrix for a decision tree.

Therefore, the complete list of learning objectives, with the verb matching the objective to one of Bloom's taxonomy categories is the following:

1. **Explain** when and why non-linear classifiers are needed (Synthesis)
2. **Explain** the basic concepts of the decision trees classifier (Comprehension)
3. **Explain** the underlying algorithm of decision trees and how they are trained (Comprehension)
4. **Implement** a decision tree (Application)
5. **Explain** why and how one can combine multiple classifiers (Synthesis)
6. **Contrast** a decision tree and a random forest (Analysis)
7. **Explain** the basic concepts of the random forests classifier. (Comprehension)
8. **Implement** a random forest classifier (Application)
9. **Describe** the information gain function (Knowledge)
10. **Describe** different node-splitting criteria such as misclassification, entropy, and Gini Index (Knowledge)
11. **Compute** the confusion matrix for a decision tree (Application)

This list will be used throughout the rest of the research in order to design effective assessments.

4 Designing effective assessments

According to "The Student Assessment Handbook" by L. Dunn et al. [11], designing effective assessments is one of the most difficult tasks for every professor when designing a course. Different assessment types can be used in order to test students' knowledge, however, each type has specific advantages and disadvantages.

The book by L. Dunn et al. introduces multiple types of assessments that can be used to test student knowledge. For the scope of this research, we will only focus on summative assessments and performance-based assessment types. The former type is generally used in order to test students' knowledge at the end of the learning period, usually through an examination at the end of the course. The latter focuses on evaluating students' ability to apply knowledge and skills to real-world tasks or scenarios. For example, students would be asked to complete an implementation assignment that simulates the need of using the decision tree classifier for a given problem. We will only focus on these particular assessment types since we believe they are the most feasible within the context of the ML course, given the relatively short amount of time students need to learn and prove their understanding. Factors such as the large number of students participating in each edition of the course and the limited number of Teaching Assistants to help with the process have also been taken into consideration when choosing to only focus on these types of assessments.

After combining the possible question types provided by both summative assessments and performance-based assessments, four types of questions [11] were extracted. Interviews with the responsible professors of the course and the existing literature on teaching Computer Science courses [3] [9] helped uncover and validate our initial presumption based on personal experience that these are the most effective types of questions that can be used in order to measure undergraduates' knowledge. Thus, each question will belong to one of the following types:

1. Multiple-choice question (Summative assessment): This type of question is used to test the lower-end of the cognitive levels. Students only need to choose the correct answer and the grading process is very time-efficient, as it can even be automated.

2. Open-ended questions (Summative assessment): This type of question is used to test students' understanding of more complex learning outcomes. Students are required to explain concepts in short sentences or paragraphs. However, this type of question can sometimes be time-inefficient.

3. Implementation exercise during the exam (Performance-based assessment): These questions are particularly great for assessing objectives corresponding to the Application cognitive layer. Grading can be automated to some extent, however, students are generally required to spend a large amount of time to complete them.

4. Homework assignments (Performance-based assessment): This assessment type can be useful since the assessment process is mitigated throughout the course, rather than only being done during the end examination period.

Throughout this research, the terms reliability and validity are used in order to describe the ”fitness” of the questions. The term ”reliability” refers to how accurate and repeatable test scores are [19]. In other words, a question is only considered reliable if different students with equal understanding levels receive a similar grade for their answers and the grade reflects their actual knowledge as accurately as possible. Furthermore, the term ”validity” refers to how meaningful, useful, and appropriate the test scores are. Conventionally, validity is defined as the amount to which a test actually measures what it is intended to assess [19].

Intuitively, one might be inclined to believe that open questions uncover students' understanding levels in a more reliable and accurate fashion. However, recent studies suggest that this belief is not necessarily true and it depends on the level of understanding that is tested. Literature [13] [2] suggests that there are no significant differences between multiple-choice questions and open-ended questions when the knowledge, comprehension, and application cognitive domain levels are tested. We also asked the opinion of TAs regarding this matter, and we found that 57% of ML TAs agree that, indeed, students generally perform equally when given either one of these question types. The survey results for this statement can be found in Appendix B and further discussions regarding the survey setup and results can be found in Sections 5 and 6. However, for open-ended questions, it is also important to take into consideration the fact that the grading process becomes more difficult and time-consuming. This is important to note because, due to the limited time and strict constraints of the grading process imposed by the TU Delft board of examiners, assessment questions should be created such that the grading process is as time-efficient as possible. Therefore, learning outcomes are ideally mapped to test questions being structured as multiple-choice questions whenever possible and they will be preferred when the description of the learning objective allows it.

Mapping the learning objective with the most efficient type of assessment question is done based on each learning objective's usage of verbs from Bloom's taxonomy. We will consider that the question type is the ”most efficient type” when the grading process cannot be done in a more time-efficient manner, without significantly reducing the reliability or validity of a question. Table 1 depicts the mapping of each learning objective to its corresponding preferred question type, according to the literature on this subject [13] [2], combined with the time-efficiency requirements described above.

Verb used	Cognitive level	Preferred question type
Explain	Comprehension	Multiple-choice
Explain	Synthesis	Open-ended
Implement	Application	Implementation / Homework
Contrast	Analysis	Open-ended
Describe	Knowledge	Multiple-choice
Compute	Application	Multiple-choice / Open-ended

Table 1: The mapping of verbs corresponding to Bloom's taxonomy cognitive levels to the preferred assessment question type.

In order to prove the validity and reliability of the mappings within the ML context, we decided to create a set of possible question formulations containing both multiple-choice and open-ended alternatives. In particular, we decided to test whether multiple-choice questions can reliably and efficiently be used for assessing students' understanding at the level of ”Application” within Bloom's taxonomy. We decided to test the fitness of multiple-choice questions for this particular cognitive level, since, the literature suggests that it can generally be reliably tested with multiple-choice questions, however, it depends on the domain of the question. This is also the reason for Table 1 listing both summative question types as the preferred type. Thus, we decided to ensure that the ”Application” cognitive level is, indeed, assessable with multiple-choice questions within the ML domain. Therefore, the set of questions has been created based on learning objective number 11.

The questions designed are all formulated as follows: *Given the following sample data and the following decision tree, what would be the resulting confusion matrix?*

1. Open-ended question: *Write down the complete confusion matrix as seen during the lectures*
2. Alternative open-ended question: *Only write down the number of false-positives.*
3. Multiple-choice question: *Pick the correct confusion matrix from the following options: ...*

Assessing the reliability of these different possible questions was done from the ML course staff's perspective (teachers and TAs). Interestingly, 57% of the survey participants believe that the multiple-choice question formulation is the best formulation out of the three. More information about the survey's setup can be found in the following section, while the results of this question can be found in Appendix B.

It is also important to analyze potential homework assignments when creating different assessment methods. According to M. Moravec et al. [10], learning before lectures can significantly improve students' performance and knowledge retention. Therefore, the introduction of such assignments would, in theory, benefit students in understanding the taught modules more effectively. These assignments would be especially useful in order to teach random forests. This is due to the fact that the random forest classifier is defined based on the decision trees classifier. Hence, students are required to understand key concepts of decision trees in order to be able to progress. Further discussions regarding this assignment proposal can be found in Sections 5 and 6.

5 Interviews and Surveys: Setup and Results

In order to gain more valuable insights regarding this topic, we initially orchestrated a series of interviews with the head teacher of the ML course. We additionally conducted a series of surveys with another set of 7 TAs in order to supplement our findings about the subject in a more time-efficient manner. In order to create the list of TAs asked to complete the survey, we searched through the public records of TAs, given by TU Delft, for the 2021-2022 and 2022-2023 editions of the course. Every person found in these records was asked to complete this survey. While data from 7 TAs could indicate interesting underlying patterns, the results should mostly be considered as indicative and further research is needed in order to finalize the results and reach a better level of understanding about the topic.

5.1 Interviews

We started the experimental work by conducting a series of interviews with the responsible professor of the course. The questions asked during the interviews can be found in Appendix A of this research. The resulting data has mainly been used to further improve the learning objectives of the course, as well as to better understand assessment requirements. However, we also received valuable feedback during this process. One of the key ideas uncovered by this step is that assessments could also be structured as assignments that students are required to complete during the course. Therefore, these assignments could be used to measure the progress of students for the decision trees module in order to be able to progress to learning random forests, since the latter module uses similar concepts and essentially expands upon the former module. This idea is consistent with the literature found on this topic [10], which has also been mentioned in the previous section. Using these insights, we designed the surveys such that TAs could select the most important concepts learned throughout the decision trees module. We refer to the ”most important” concepts as the concepts which are later used during the random forests module and that usually cause the most difficulties for students.

5.2 Surveys

Surveys have been used to further improve our understanding of assessment creation and best practices, while also increasing the efficiency and decreasing the time required by participants to answer the questions, as opposed to organizing interviews with them. Although the nature of surveys restricts us from asking follow-up questions for interesting ideas, we were able to obtain a relatively significant amount of participants willing to answer the questions. We also included feedback sections where participants were allowed to share other thoughts or ideas in order to minimize the negative effects bound to surveys. The survey questions can be found in Appendix B, together with the image containing the complete overview of the visualizations for the obtained data, as produced by the Google Forms website.

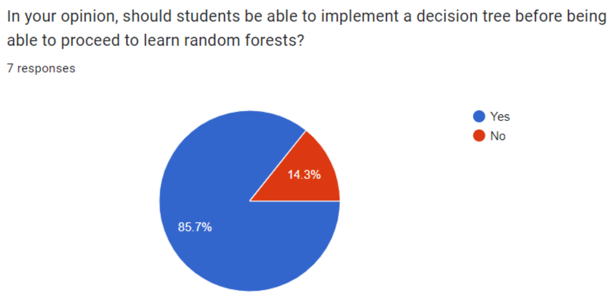
The participants of the survey were selected such that they meet the following predefined requirements:

- Participants need to be enrolled or have graduated the Computer Science & Engineering Bachelor's degree from TU Delft.
- Participants are required to have previous experience with being a TA for the ML course.
- Participants are required to have previous experience with grading exams or assignments for the ML course.

It is important to ensure that participants meet these requirements, therefore, they were specifically selected from the online records of Machine Learning TAs posted by TU Delft.

Participants were also asked if they meet these requirements during the survey, and responses that did not meet the criteria have been removed from the final results.

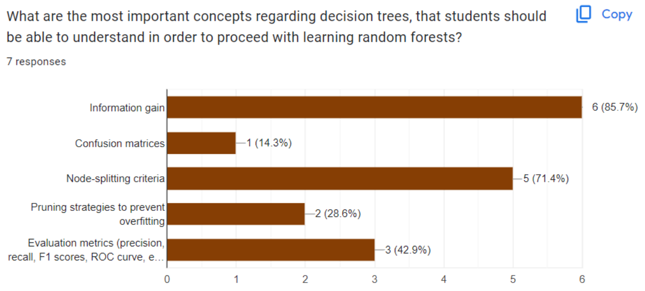
In order to test the veracity of the interview findings, participants were asked their opinion regarding the need for students to fully understand the key concepts of decision trees before being able to proceed to learn random forests. Figure 2 depicts the results for this question. It is interesting to note that approximately 86% of participants believe that it is, indeed, better if students' understanding of the decision tree classifier is tested before being able to continue with the latter.


In your opinion, should students be able to implement a decision tree before being able to proceed to learn random forests?

7 responses

Figure 2: The visualization of survey results for the question *"Should students be able to implement decision trees before being able to proceed to learn random forests"*.

Naturally, there are certain concepts within decision trees that are more important than others. Hence, the participants were asked an additional question in order to understand which concepts are more relevant. A wide majority of participants (85%) believe that understanding the information gain function is of key importance when learning the decision tree classifier. Node-splitting criteria and evaluation metrics also appear to be important, with 71% and 42% of participants selecting them, respectively.


What are the most important concepts regarding decision trees, that students should be able to understand in order to proceed with learning random forests?

7 responses

Figure 3: The most important concepts within decision trees.

These results help us better pinpoint what students should be able to understand before "overwhelming" them with additional information, such as random forests. In the following section, we further discuss how these results can be used to create a homework assessment that should be used to test students' understanding before teaching them random forests.

6 Discussion

Taking into account all of the data mentioned above, we propose the following addition to the course; an assignment in which students are asked to implement parts of a decision tree, such as node-splitting criteria and the information gain function. Only after successfully completing the minimum amount of points in this assignment should they be allowed to start learning and completing the afferent random forests assignments and learning objectives. This course design has been seen before in other courses within the Computer Science curriculum of TU Delft, where assignments were mandatory and "unlocked" based on a predefined order of importance that aimed at maximizing students' understanding of particular topics (for example, the Concepts of Programming Languages course, during the 2021-2022 edition). Figure 3 indicates that TAs believe that certain parts of the assignment are less important than others, such as confusion matrices, hence we propose that these parts should be given as "boilerplate code" within the assignment. This should theoretically help students complete the assignment faster, while still learning these concepts by reading the given code and trying to use it in order to create the functions for the concepts deemed as more important. Of course, this requires further research in order to get a better understanding of whether the hypothesis actually holds. This proposal would also be consistent with the feedback we received from the survey, where TAs noted that the different node-splitting criteria are more important to learn as concepts and their possible applications, rather than purely as formulas. This feedback can also be found in Appendix B. We suspect that, by creating this assignment, students would see a possible application and understand the concept better. However, further studies are, again, required in order to prove this. This is also discussed in the Future Work section of the paper.

In conclusion, we propose the following assignment formulation: Students should be asked to complete a Python program that implements decision trees. In this assignment, they would be given already-implemented functions that calculate the confusion matrices and pruning strategies to prevent overfitting, while being asked to implement the information gain function and different node splitting criteria. Essentially, students would be asked to create an algorithm that is efficient enough at predicting data, given predefined training and test data sets. By comparing students' algorithm predictions to an optimal implementation of the decision tree classifier, teachers would theoretically be able to approximate student solutions' closeness to the optimal algorithm, hence grading them accordingly.

7 Responsible Research

When analyzing the response data from interviews and surveys, it is important to consider the ethical implications of the research and discuss how the research can be conducted in a responsible manner. This includes ensuring that the data collected is accurate and unbiased and that the privacy of the teachers and TAs is protected.

7.1 Ethical considerations

Responsible research in assessing students' understanding of ML concepts involves addressing ethical considerations. It is essential to obtain informed consent from participants and ensure their privacy and confidentiality. Teachers and TAs should be made aware of the purpose, risks, and benefits of the interview and survey process. Additionally, efforts should be made to minimize potential biases and ensure that no harm is caused to the participants.

7.2 Data Privacy and Security

In collecting and analyzing the interviews and surveys, data privacy and security must be safeguarded. Collecting and storing response data should adhere to established privacy regulations and institutional policies. Anonymization and de-identification techniques have been employed to protect individual personnel identities. In this sense, survey participants were not asked to complete their names since the forms provided by Google Forms are not believed to be particularly safe with regard to data security. While the results aggregation for surveys is provided in Appendix B, individual response registrations will not be presented in order to minimize potential identification based on the list of responses. Access to individual response data is restricted to authorized personnel only, and appropriate measures have been implemented to prevent data breaches. As such, the individual anonymized response data will only be made available for further inspection and analysis upon request.

7.3 Fairness and Bias Mitigation

Responsible research in assessing response data requires addressing issues of fairness and bias. Results must be unbiased and free from any discriminatory factors such as gender, race, or socioeconomic background. Steps should be taken to ensure that the response data and scoring criteria do not disadvantage any particular group of TAs.

8 Future work

The main goal of this research was to provide guidelines and review the assessment techniques used to measure students' progress for the decision trees module within the ML course of TU Delft. Due to the very limited time constraints of this research, further work is recommended in order to deeper analyze certain aspects of the course.

First, different modules must be further analyzed in order to improve the course as a whole. While the improvements proposed by this research can improve the decision trees and random forests modules and there is previous literature [20] which aims to improve the non-parametric classifiers, it is not yet sufficient in order to completely restructure the course based on the provided guidelines. Modules such as linear classifiers, perceptrons, etc. need further studying and their corresponding learning objectives and assessments require further independent reviewing and analysis.

Although this research used different questioning techniques that aim at revealing professionals' opinions such as the teaching staff of the course, the collected data only represents a fraction of the possible data. As such, a higher number of teachers and TAs should be interviewed in order to gain better knowledge and possibly more accurate and complete results regarding this subject.

Yet another suggestion for further research on this subject would be to also take into consideration the students' opinions regarding this matter. We believe that it could be possible that students might perform better if the course would focus more on designing homework assignments that can accurately reflect student knowledge without the very high time pressure students face within the exam period. However, this hypothesis needs further studying, and controlled groups of students should be tested in order to complete such experiments.

9 Conclusions

This research provides an extensive analysis of the design of assessment questions that accurately measure student progress within the decision trees and random forests classifiers modules of the ML course.

Learning outcomes were reviewed and extended upon such that they better reflect teachers' expectations regarding student knowledge. The process is based on Bloom's taxonomy of learning, which provides a framework for analyzing the cognitive levels of understanding. Each learning objective was individually matched with its corresponding level of understanding in order to gain a better overview and be able to provide more valid and reliable assessments.

Based on the cognitive level of understanding assigned to each learning objective, a series of possible assessment questions and guidelines for creating effective questions were devised. The proposals are consistent with existing literature on similar subjects.

Further insights from professionals such as teachers and TAs within the teaching staff of the ML course have been compiled, analyzed, and discussed during this research. The aforementioned professionals have been interviewed and asked to complete a survey designed to unveil a series of assessment requirements, expectations, and experience-based opinions. Based on the unveiled information gained by this process, we proposed a set of alternative assessments, including a homework assignment that can be used to assess student knowledge regarding only the decision tree module, rather than the complete course.

Altogether, this research provides a replicable process of analyzing, improving, and extending the predefined learning objectives of the decision trees and random forests modules using Bloom's taxonomy of learning. This research also proposes a series of viable multiple-choice questions that can be used as a template for questions created for future exams. Last but not least, we provide valuable insights from professionals in the field of teaching ML to undergraduate students and give an alternative assessment method that focuses on measuring students' progress during the course, rather than only during the final examination period.

A Interview questions

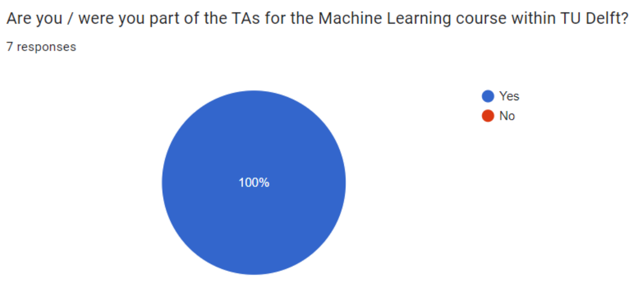
The following questions were asked during the interviews with both the course coordinator and TAs:

1. How do you create assessment questions? How does the process look like?
2. Are there any predefined mappings that match the learning outcomes' cognitive levels to the type of question? For example, if the learning outcome is ”Contrast a decision tree classifier with a random forest” do you automatically create an open question since the verb contrast requires a deep level of understanding or do you attempt to create a multiple-choice question, and if unsuccessful you fallback to the open-ended type?
3. Which of the following questions do you think would reflect the students' understanding levels better?
   - Question 1 (Multiple choice): Choose the correct formula for the entropy node-splitting criteria:
     (a) $-\sum_c p_c \log(p_c)$
     (b) $\max_c(1-p_c)$
     (c) $\sum_c p_c(1-p_c)$
     (d) $\sum_c p_c \log(p_c)$
   - Question 2 (Open question): Explain in one sentece what node-splitting means and write the formula for entropy.
4. How good is the question chosen above to assess understanding of the following learning objective: ”Describe different node-splitting criteria such as: missclassification, entropy and Gini Index?” (Likert scale)
5. Why did you rate the question this much? What is the rating based on? Any feedback?
6. What type of question would you choose for this learning outcome: ”Compute and analyze the confusion matrix of a decision tree”?

   The question would be structured like this: ”Given the following sample data and the following decision tree, what would be the resulting confusion matrix?”

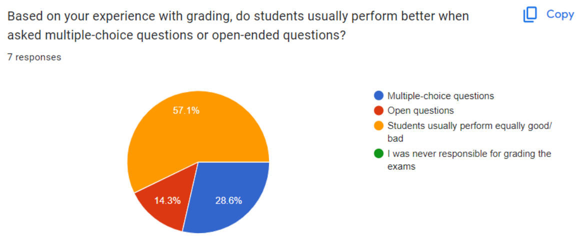
   Possible question options could be something like:
   - Pick the correct answer (Multiple-choice question).
   - Compute the number of true-positives. (open question - partial testing of understanding)
   - Compute the full confusion matrix. (open question - testing full understanding)
7. Based on your experience with grading, do students usually perform better when asked multiple-choice questions or open-ended questions? (Possible answers are: Multiple-choice, Open-ended, no noticeable differences, they perform equally good/bad)
8. Are you familiar with Bloom's taxonomy? If so, would you say that learning objective ”Explain when and why non-linear classifiers are needed” rather has the depth of 'understanding' or 'synthesis'?

B Survey questions



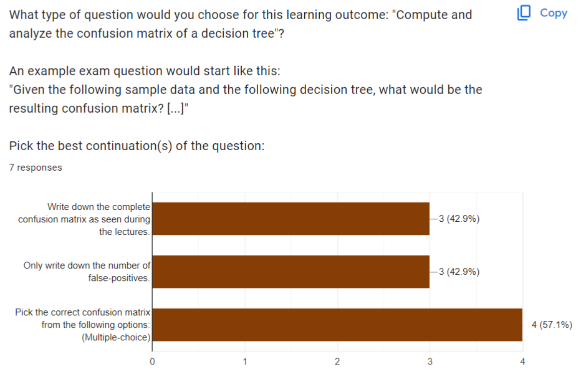
Are you / were you part of the TAs for the Machine Learning course within TU Delft?

7 responses



Based on your experience with grading, do students usually perform better when asked multiple-choice questions or open-ended questions?

7 responses

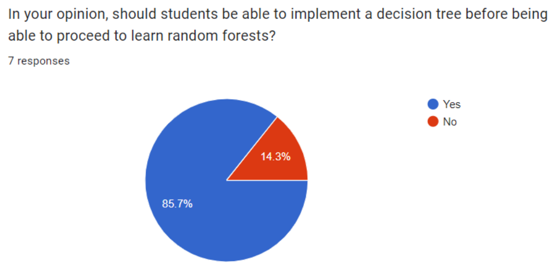


What type of question would you choose for this learning outcome: "Compute and analyze the confusion matrix of a decision tree"?

An example exam question would start like this:
"Given the following sample data and the following decision tree, what would be the resulting confusion matrix? [...]"

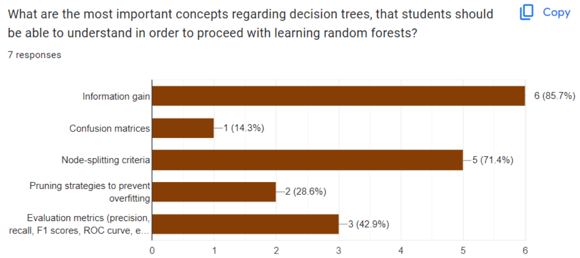
Pick the best continuation(s) of the question:

7 responses



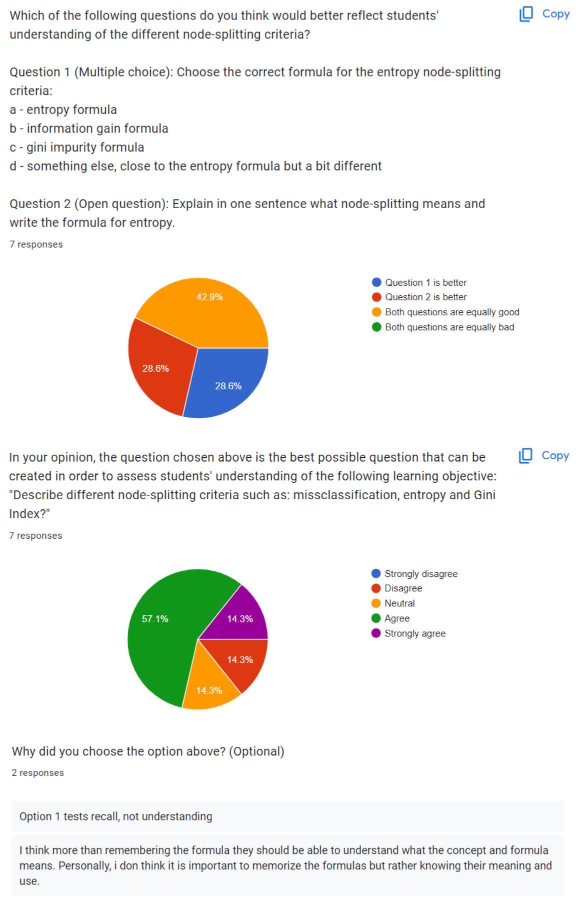
In your opinion, should students be able to implement a decision tree before being able to proceed to learn random forests?

7 responses



What are the most important concepts regarding decision trees, that students should be able to understand in order to proceed with learning random forests?

7 responses



Which of the following questions do you think would better reflect students' understanding of the different node-splitting criteria?

Question 1 (Multiple choice): Choose the correct formula for the entropy node-splitting criteria:
a - entropy formula
b - information gain formula
c - gini impurity formula
d - something else, close to the entropy formula but a bit different

Question 2 (Open question): Explain in one sentence what node-splitting means and write the formula for entropy.

7 responses

42.9%
28.6%
28.6%
● Question 1 is better
● Question 2 is better
● Both questions are equally good
● Both questions are equally bad
Copy

In your opinion, the question chosen above is the best possible question that can be created in order to assess students' understanding of the following learning objective: "Describe different node-splitting criteria such as: missclassification, entropy and Gini Index?"

7 responses

Why did you choose the option above? (Optional)

2 responses

Option 1 tests recall, not understanding

I think more than remembering the formula they should be able to understand what the concept and formula means. Personally, i don think it is important to memorize the formulas but rather knowing their meaning and use.

References

[1] Benjamin S Bloom and David R Krathwohl. *Taxonomy of educational objectives: The classification of educational goals. Book 1, Cognitive domain*. longman, 2020.

[2] Brent Bridgeman and Donald A Rock. Relationships among multiple-choice and open-ended analytical questions. 30(4):313–329, Dec 1993.

[3] Qingwan Cheng, Angela Tao, Huangliang Chen, and Maira Marques Samary. Design an assessment for an introductory computer science course: A systematic literature review. pages 1–8, 2022.

[4] Wikipedia Contributors. Bloom’s taxonomy. https://en.wikipedia.org/wiki/Bloom’s_taxonomy, Mar 2023. Accessed: 2023-05-15.

[5] Johannes Fürnkranz. Decision tree. pages 263–267, 2010.

[6] Christiane Gresse von Wangenheim, Nathalia da Cruz Alves, Marcelo Rauber, Jean Hauck, and Ibrahim Yeter. A proposal for performance-based assessment of the learning of machine learning concepts and practices in k-12. *Informatics in Education*, 09 2022.

[7] IBM. Ibm global ai adoption index 2022 new research commissioned by ibm in partnership with morning consult. 2022.

[8] Declan Kennedy. *Writing and Using Learning Outcomes*. 2007.

[9] Andrew Luxton-Reilly, Jacqueline Whalley, Brett Becker, Cao Yingjun, Roger McDermott, Claudio Mirolo, Andreas Mühling, Andrew Petersen, Kate Sanders, and Simon. Developing assessments to determine mastery of programming fundamentals. pages 47–69, 01 2017.

[10] Marin Moravec, Adrienne Williams, Nancy Aguilar-Roca, and Diane K. O’Dowd. Learn before lecture: A strategy that improves learning outcomes in a large introductory biology class. *CBE—Life Sciences Education*, 9(4):473–481, 2010. PMID: 21123694.

[11] Chris Morgan, Lee Dunn, Sharon Parry, and Meg O’Reilly. *The Student Assessment Handbook*. Dec 2003.

[12] Geraldine O’Neill and Feargal Murphy. Guide to taxonomies of learning. 2010.

[13] Murat Polat. Analysis of multiple-choice versus open-ended questions in language tests according to different cognitive domain levels analysis of multiple-choice versus open-ended questions in language tests according to different cognitive domain levels polat 77. 2020(2):76–96.

[14] S.R. Safavian and David A Landgrebe. A survey of decision tree classifier methodology. 21(3):660–674, Jan 1991.

[15] Yan-Yan Song and Ying Lu. Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, 27(2):130–5, 2015.

[16] Christopher Starr, Bill Manaris, and Roxann Stalvey. Bloom’s taxonomy revisited: Specifying assessable learning objectives in computer science. *ACM Sigcse Bulletin*, 40:261–265, 01 2008.

[17] Elisabeth Sulmont, Elizabeth Patitsas, and Jeremy R. Cooperstock. What is hard about teaching machine learning to non-majors? insights from classifying instructors’ learning goals. *ACM Trans. Comput. Educ.*, 19(4), jul 2019.

[18] Jonathan Supovitz. Getting at student understanding - the key to teachers’ use of test data. *Teachers College Record*, 2016.

[19] Nathan A Thompson. Reliability & validity. *Assessment Systems*, 2013.

[20] Kerem Çakıcı. A guidline for creating assessments in machine learning education. *Tudelft.nl*, 2022.