GRADE 6 STUDENTS' PERCEPTION AND USE OF DATA-BASED DECISION TREES

Susanne Podworny, Yannik Fleischer, and Sven Hüsing
Paderborn University, Paderborn, Germany
Podworny@math.upb.de

Decision-making processes are often based on data and data-driven machine learning methods in different areas such as recommender systems, medicine, criminalistics, etc. Well-informed citizens need at least a minimal understanding and critical reflection of corresponding data-driven machine learning methods. Decision trees are a method that can foster a preformal understanding of machine learning. We developed an exploratory teaching unit introducing decision trees in grade 6 along the question "How can Artificial Intelligence help us decide whether food is rather recommendable or not?" Students' performances in an assessment task and self-assessment show that young learners can use a decision tree to classify new items and that they found the corresponding teaching unit informative.

INTRODUCTION

Is popcorn more recommendable than an apple? What about strawberry yogurt? The topic of recommended diets is frequently addressed in the media, for example in an article in the New York Times (Quealy & Sanger-Katz, 2016). In many countries, there is mandatory nutrition labeling based on the "big seven": fat, saturated fatty acids, carbohydrates, sugars, protein, energy, and salt. In Germany, this nutritional data can be found on nearly every food product. Is it self-evident to say that a product that contains little fat is more likely to be recommended? No, more than one variable should be considered. A combination of different characteristics can be used to make predictions about whether a particular food item is more likely to be recommendable or not.

Artificial intelligence (AI) and machine learning (ML) methods are increasingly permeating general life. This is accompanied by the call to teach data science and AI content early in school (Biehler & Schulte, 2018; Engel, 2017). In more recent approaches, there is a focus on AI and ML (Long & Magerko, 2020). The recently published German Data Literacy Charter also emphasizes the importance of critical and competent handling of data and data-based decision-making processes:

> In concrete terms, this requires the inclusion of data literacy in the curricula and educational standards of schools […]. Learners should not only be addressed as passive consumers of data. We rather want to enable them to actively shape data-related insights and decision-making. (Schüller et al., 2021, p. 3)

Introducing the idea of decision trees with an unplugged activity is the goal of a teaching unit we developed for grade 6 students (aged 11–12) in the ProDaBi project (www.prodabi.de/en) (Podworny et al., 2021). Students use data cards about 55 food items with data on nutritional values to develop decision rules that classify food items as "rather recommendable" or not. After these hands-on experiences, it is discussed how computers may take over an automatic creation of decision trees. Several classes have participated in the teaching unit. In this paper, we examine (a) the students' perception of the teaching unit and (b) how students use and reason about data-based decision trees when classifying new food items.

BACKGROUND

Decision Trees in Machine Learning

AI and ML is a broad field with different methods and algorithms, but all are data-driven (Hastie et al., 2009). A distinction is made between supervised and unsupervised learning (Hastie et al., 2009). Unsupervised learning is about finding patterns in data; supervised learning is about explaining patterns in data and making predictions. Classification is a supervised learning problem that aims for assigning objects described by variables to a particular class (for example, predicting the presence of a disease by diagnostic features). Thus, we look for a classifier that uses predictor variables to predict the value of a target variable (class). A decision tree that is shaped as a hierarchical tree structure of decision rules can be used as such a classifier and can be algorithmically constructed from data (Breiman et al., 1998). If the tree structure is not too large, the decision-making is transparent and easy to interpret (Engel et al., 2018) and therefore well suited for teaching (Martignon et al., 2003). Thus, decision processes can be traced and patterns in the data can be analyzed. Another advantage is that higher mathematics is not required for understanding decision trees. Machine learning with decision trees is also a substantial part of the International Data Science in Schools Project (IDSSP Curriculum Team, 2019).

Teaching Decision Trees

The research field on the understanding of machine learning and related concepts is still young and diverse. In a study by Sulmont et al. (2019), university students without much prior knowledge of mathematics or computer science gained an understanding of machine learning algorithms, but had difficulty developing models themselves. Zieffler et al. (2021), on the other hand, found that secondary school teachers were able to create and read decision trees but had difficulty evaluating the models. The common finding is that adult learners could be taught the decision tree method without much prior knowledge. We now want to use our study to investigate what this looks like with young learners.

A study examining 10–13 year old students was able to show that these students developed a basic understanding of machine learning concepts and were able to apply them in other situations (Hitron et al., 2019). Hitron et al. (2019) divided the learning process into six steps: (a) data labeling; (b) feature extraction; (c) model selection; (d) parameter tuning; (e) evaluation; and (6) real-world application. Asking which of these steps can be taught explicitly and which can be taught as a black box, it was shown that understanding machine learning processes was only generated when learners at least had their own experiences labeling data (a) and evaluating the model (e). Hitron et al. (2019) encourage further research whether other steps of the process can be uncovered for students to improve their understanding.

RESEARCH QUESTION

The results of Hitron et al. (2019) are promising that even young learners can understand and apply machine learning concepts. However, in our view, an important part of teaching machine learning is teaching about

- what a possible model looks like,
- giving students an idea of how a model can be built from data, and
- why it might work the way it does.

Leaving the building process of a model as a black box might give the impression that it is a mystic and incomprehensible process. We would rather argue for teaching the underlying ideas of deriving a model from data to demystify the process. In particular, creating and applying a model is the focus of our teaching unit.

In this study, we first focus on students' understanding of a decision tree model as a classifier. Additionally, we want to assess the students' attitudes towards the topic of data-based decision trees and the context of nutrition. Therefore, we pose two research questions:

- How do students like the teaching module about data-based decision trees for classifying food items?
- How do students use and reason about data-based decision trees when classifying new food items?

METHOD

We developed a series of nine lessons and conducted them in several classes in grade 6. The main approach of our teaching unit is described in Podworny et al. (2021).

Study

After the students attended the series of lessons, each of them answered a questionnaire with two parts. Students were asked in a first part about their personal experiences and attitudes concerning the series of lessons. Among other questions, they were asked questions regarding the extent to which they liked the teaching module (5-point Likert scale) and the context of food (5-point Likert scale) and whether or not they recommend the teaching unit (yes, no, don't know) for other students.

In the second part, questions were asked about applying a decision tree. The students were asked whether they can explain how a decision tree works (yes, no, don't know). To not just ask, but also test whether the students understand how a decision tree works, the students executed different tasks to show the understanding they gained. One task was designed to assess whether the students can use a given decision tree to classify a new food item.

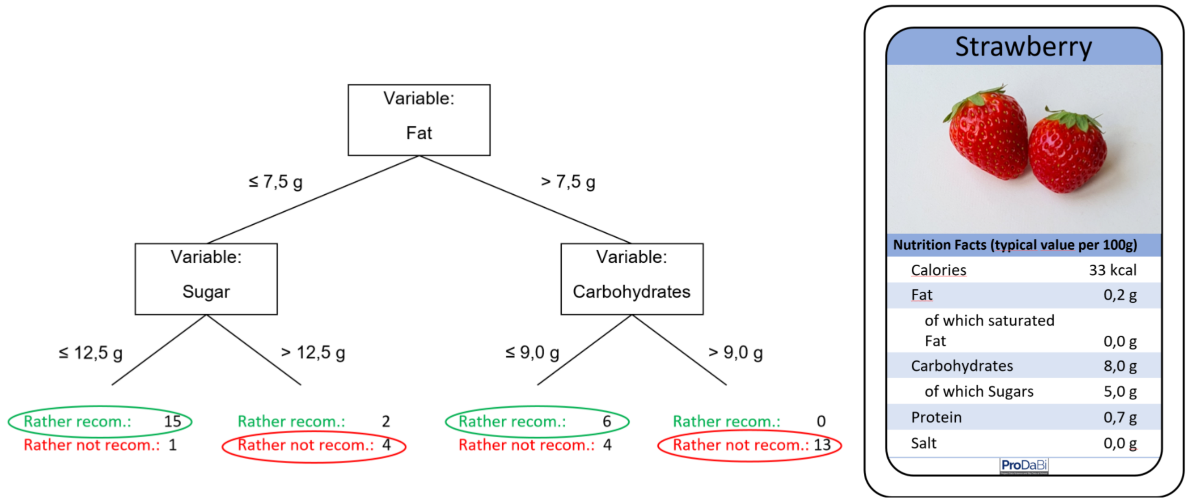

Figure 1. Given decision tree (left) and a new data card for strawberries (right)

Figure 1 shows the given decision tree and a data card for strawberry as new food item that is to be classified. The representation of both the decision tree and the food item were chosen in a way that students knew from class. The first task was "*Here you can see a decision tree created with the computer and a new food card. Use the decision tree to decide whether strawberries are rather recommendable or rather not recommendable.*" Following the decision tree, the correct answer is "rather recommendable." The second task was "*How did you decide on the classification 'rather recommendable' or 'rather not recommendable' with the help of the decision tree? Finish formulating the decision rule. The strawberry is classified as … because … .*" An ideal student solution might look like this: Strawberry is rather recommendable because it has 0.2 g fat and 5.0 g sugar per 100 g. That is less than the threshold of 7.5 g fat and less than the threshold of 12.5 g sugar. The majority of food items in the respective path of the decision tree are rather recommendable.

Data and Participants

Children from nine different classes and four different schools answered the questionnaire. Of those, 122 are female and 111 are male, a total of 233 children. Most of them are 11 or 12 years old. All of them participated in the teaching unit, taught by five different computer science teachers. None of the students had a background in statistics, and they had between a half and a full year of computer science classes, but without any ML or AI topics. The teachers got all the teaching material (lessons overview and description, worksheets, presentations) and a brief introduction by the authors in advance and carried out the lessons autonomously.

Answering the questionnaire took place in the classes, several lessons after the teaching unit, either as a pen-and-pencil questionnaire (n = 83) or online (n = 150). Both implementations were equivalent.

Analysis Method

The items concerning attitudes and self-assessment are evaluated quantitatively. Qualitative content analysis (Mayring, 2015) is used for evaluating the second task of part two of the questionnaire about students' explanations. The aim of this systematic and rule-guided analysis is to identify structures and summarize students' explanations. The coding unit for the analysis is a meaning section of a student's statement. The statements were coded independently by the three authors of this paper. In case of unequal coding, discussion was held until agreement was reached.

RESULTS & DISCUSSION

We start with results concerning the students' attitudes towards the teaching unit (first research question). Figure 2 shows results for the item "How well did you like working with the topic of food?" (item context food) and "Overall, how well did you like the series of lessons on decision trees?" (item teaching module).

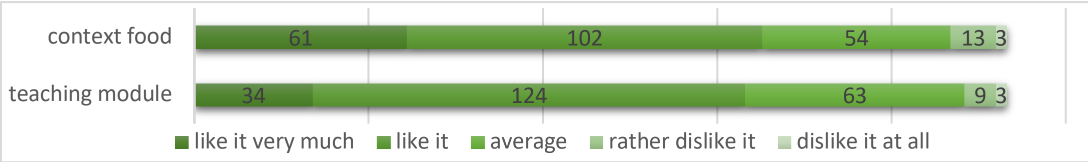

Figure 2. Results for students' attitudes towards the teaching module and the food context (n = 233)

Summarizing the results for the first item on the food context for "like it very much" and "like it," there is a positive agreement of 163 children (70%) and a disagreement "rather dislike it" and "dislike at all" of only 16 children (7%). Fifty-four children (23%) are neutral about the food context. The teaching module was rated positively by 158 children (68%), negatively by 12 children (5%), and neutral by 63 children (27%). The results show that both the lesson series and the context of food are well received by the learners. Only 7% are negative.

Additionally, students rated the item "Would you recommend the lesson series to friends from other classes?" with results shown in Figure 3. This is still a positive result with only 21 students (9) indicating rejection.

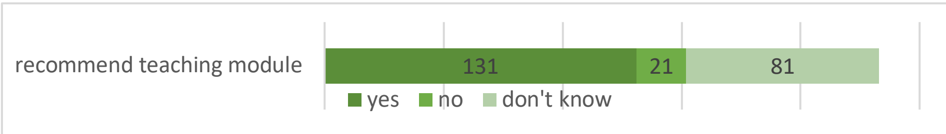

Figure 3. Results for students' recommendation of the teaching module to friends

Next, students were asked "Can you explain how a decision tree works?" (Figure 4). Most students, 196 (84), answered yes, 29 (12) said no and 8 (3) were not sure. The vast majority were sure regarding their understanding. The next item from part two of the survey showed that this was justified for the simple application of a decision tree and a new food item (Figure 1).

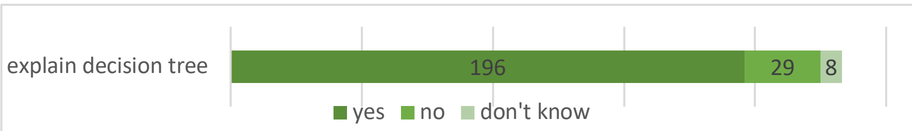

Figure 4. Results for student's self-assessment on explaining a decision tree

The strawberry was identified correctly as rather recommendable by 223 (96) children (Figure 5). This is more children than previously indicated to be able to use a decision tree. Therefore, it is valuable to analyze the written explanations in which reasons for the classification of the strawberry can be found.

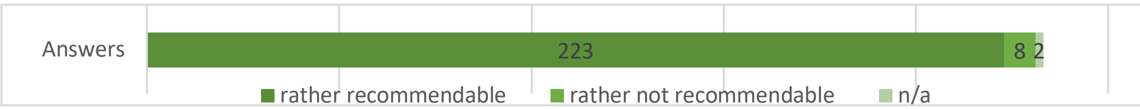

Figure 5. Results for students classifying the strawberry with a given decision tree

Going to the second part of the analysis in rating the explanations for the answers in Figure 5, the claimed knowledge appears somewhat more fragile.

An ideal student solution, which makes concrete reference to the values for fat and sugar and the two associated thresholds, was provided by 66 (28) children. A typical example was student 74 with the explanation "The strawberry is classified as recommendable because it has a value below 7.5 for the variable fat and below 12.5 for the variable sugar." Fifteen (6) other children correctly described the decision process without referring to the concrete situation from Figure 1. An exemplary description was “You have to look from the first variable (in this case fat) if it is below or above the threshold. In the next branching the same. The strawberry is rather recommendable” (student 87). Together there were 81 students (35) who gave a completely correct solution.

Further descriptions could be summarized by students who referred either generally to the characteristics of fat and sugar (21 children, 9) or generally to the thresholds (19 children, 8). Example explanations were “The strawberry is classified as recommendable because it is always below the threshold value” (student 9), or “I looked how much fat the strawberry has and then classified. Next, I classified sugar and then that was recommendable.” We rated these as partially correct answers that contained at least essential components of the correct solution. It could be assumed that these 40 (17) children understood the process but just did not write it down in detail.

Several children (25; 11) referred to a decision-making process in very general terms, with no specific reference to the given tree or strawberry data. Some explanations showed a minimum reference to the concrete situation such as student 85 “I looked at the object and decided with the decision tree whether it is rather recommendable or not.” Others wrote even more generally, such as student 6 “The strawberry is classified as recommendable because the AI has classified the food and thus checked each variable.” When interpreting, it was unclear if these learners understood how to use the tree because the explanations were not detailed enough.

Other explanations did not show an understanding of the classifying process at all. Thirty (13) children used their general knowledge or experience to justify a decision about strawberries. Student 55 was one example “The strawberry is classified as recommendable because they are healthy and do not have a lot of fat.”

About one fifth of the children (49; 21) gave no explanation or an inappropriate explanation such as student 127 “I got on very well with it. I did it step by step” or student 160 “I looked at the numbers.” An interpretation concerning their understanding of the process was hardly possible from these explanations.

Finally, there were eight children who partly misinterpreted the decision tree because they used the two variables, fat and carbohydrates, instead of fat and sugar. They used the wrong branch in the first place. An interpretation for this error could possibly be found in student 88’s explanation “I first looked to see if the strawberry had more than 7.5 fat, it has less. Then I looked to see if it had more than 9.0 carbohydrates. It has less. That is why it is a recommendable food.” Maybe this student did not read the mathematical signs $\leq$ and $>$ correctly and therefore used the wrong branch, so the student reasoned correctly if we interpret a continuation error from mistaking the signs. This error has already been observed during the lessons.

SUMMARY & CONCLUSION

Our results show that using a decision tree to classify a new item can be mastered by young learners. Most students (84) were confident in their perceived ability to explain a decision tree correctly and 223 (95) classified the new food item strawberry correctly. Challenges in writing accurate explanations for the classification process emerged in some places. Overall, we rate the learners’ written descriptions as good. Eighty-one students (35) gave completely correct descriptions, 40 students (17) gave a partially correct answer, and the reasoning of 74 students (32) in written form was not documented in a way to interpret either missing or existing understanding. Thirty students (13) gave arguments from general knowledge about food, and 8 students (3) used the decision tree incorrectly. We conclude that generally, the students were able to use a decision tree for classifying a new item, and furthermore, a majority adequately reasoned about it in written form.

Furthermore, we found only a small minority of students disliked the teaching module (5%) or the food context (7) and that most of the students (56) would recommend the module for other students. This analysis is a promising start to evaluating more aspects of such a teaching module. As stated above, we are interested in assessing whether students understand (a) what a possible model looks like, (b) how a model can be built from data, and (c) why it might work the way it does. The first of these aspects was investigated in this paper and in future work we want to analyze further questions of the questionnaire and additional interviews, to assess students’ understanding of other aspects.

Referring to Hitron et al. (2019), we now have new evidence that uncovering the model building process is also possible (at least for decision trees) and desirable when teaching machine learning to young students with the goal of making machine learning processes transparent to them. Furthermore, we argue to not only frame the part of model building as model selection and parameter tuning like Hitron et al. (2019), which seems to be derived from an ML user perspective but argue to also give students insights in the ML developer perspective. We consider a look into the black box of decision trees in a didactically reduced form with young learners aged 11–12 possible.

REFERENCES

Biehler, R., & Schulte, C. (2018). Perspectives for an interdisciplinary data science curriculum in German secondary schools. In R. Biehler, L. Budde, D. Frischemeier, B. Heinemann, S. Podworny, C. Schulte, & T. Wassong (Eds.), *Paderborn symposium on data science education at school level 2017: The collected extended abstracts* (pp. 2–14). Universitätsbibliothek Paderborn.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1998). *Classification and regression trees*. Chapman & Hall.

Engel, J. (2017). Statistical literacy for active citizenship: A call for data science education. *Statistics Education Research Journal, 16*(1), 44–49. https://doi.org/10.52041/serj.v16i1.213

Engel, J., Erickson, T., & Martignon, L. (2018). Teaching and learning about tree-based methods for exploratory data analysis. In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018) Kyoto, Japan.* ISI/IASE. https://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_7F3.pdf

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Springer.

Hitron, T., Orlev, Y., Wald, I., Shamir, A., Erel, H., & Zuckerman, O. (2019). Can children understand machine learning concepts? The effect of uncovering black boxes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–11). Association for Computing Machinery. https://doi.org/10.1145/3290605.3300645

IDSSP Curriculum Team. (2019). *Curriculum frameworks for introductory data science*. The International Data Science in Schools Project. http://www.idssp.org/files/IDSSP_Data_Science_Curriculum_Frameworks_for_Schools_Edition_1.0.pdf

Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–16), Association for Computing Machinery. https://doi.org/10.1145/3313831.3376727

Martignon, L., Vitouch, O., Takezawa, M., & Forster, M. R. (2003). Naive and yet enlightened: From natural frequencies to fast and frugal decision trees. In D. Hardman & L. Macchi (Eds.), *Psychological perspectives on reasoning, judgment and decision making* (pp. 189–211). John Wiley & Sons.

Mayring, P. (2015). Qualitative content analysis: Theoretical background and procedures. In A. Bikner-Ahsbahs, C. Knipping, & N. Presmeg (Eds.), *Approaches to qualitative research in mathematics education* (pp. 365–380). Springer. https://doi.org/10.1007/978-94-017-9181-6_13

Podworny, S., Fleischer, Y., Hüsing, S., Biehler, R., Frischemeier, D., Höper, L., & Schulte, C. (2021). Using data cards for teaching data based decision trees in middle school. In O. Seppälä & A. Petersen (Eds.), *21st Koli Calling International Conference on Computing Education Research* (pp. 1–3). Association for Computing Machinery. https://doi.org/10.1145/3488042.3489966

Quealy, K., & Sanger-Katz, M. (2016, July 5). *Is sushi healthy? What about Granola? Where Americans and nutritionists disagree*. New York Times. https://www.nytimes.com/interactive/2016/07/05/upshot/is-sushi-healthy-what-about-granola-where-americans-and-nutritionists-disagree.html

Schüller, K., Koch, H., & Rampeld, F. (2021). *Data literacy charter*. Stifterverband.

Sulmont, E., Patitsas, E., & Cooperstock, J. R. (2019). What is hard about teaching machine learning to non-majors? Insights from classifying instructors' learning goals. *ACM Transactions on Computing Education, 19*(4), 1–16.https://doi.org/10.1145/3336124

Zieffler, A., Justice, N., delMas, R., & Huberty, M. D. (2021). The use of algorithmic models to develop secondary teachers' understanding of the statistical modeling process. *Journal of Statistics and Data Science Education, 29*(1), 131–147. https://doi.org/10.1080/26939169.2021.1900759