frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 21 July 2025
DOI 10.3389/feduc.2025.1595209

OPEN ACCESS

EDITED BY
Dazhi Yang,
Boise State University, United States

REVIEWED BY
Anna Wróblewska,
Warsaw University of Technology, Poland
Salim Belouettar,
Luxembourg Institute of Science and Technology (LIST), Luxembourg

*CORRESPONDENCE
Ioannis Papantonis
✉ i.papantonis@ed.ac.uk

RECEIVED 17 March 2025
ACCEPTED 23 June 2025
PUBLISHED 21 July 2025

CITATION
Bueff A, Papantonis I, Simkute A and Belle V (2025) Explainability in machine learning: a pedagogical perspective.
Front. Educ. 10:1595209.
doi: 10.3389/feduc.2025.1595209

Explainability in machine learning: a pedagogical perspective

Andreas Bueff[1], Ioannis Papantonis[2]*, Auste Simkute[3] and Vaishak Belle[2]

[1]Department of Computer and Information Science, Linköping University, Linköping, Sweden, [2]School of Informatics, University of Edinburgh, Edinburgh, United Kingdom, [3]School of Design, University of Edinburgh, Edinburgh, United Kingdom

Introduction: Machine learning courses usually focus on getting students prepared to apply various models in real-world settings, but much less attention is given to teaching students the various techniques to explain a model's decision-making process. This gap is particularly concerning given the increasing deployment of AI systems in high-stakes domains where interpretability is crucial for trust, regulatory compliance, and ethical decision-making. Despite the growing importance of explainable AI (XAI) in professional practice, systematic pedagogical approaches for teaching these techniques remain underdeveloped.

Method: In an attempt to fill this gap, we provide a pedagogical perspective on how to structure a course to better impart knowledge to students and researchers in machine learning about when and how to implement various explainability techniques. We developed a comprehensive XAI course, focused on the conceptual characteristics of the different explanation types. Moreover, the course featured four structured workbooks focused on implementation, culminating in a final project requiring students to apply multiple XAI techniques to convince stakeholders about model decisions.

Results: Course evaluation using a modified Course Experience Questionnaire (CEQ) from 16 MSc students revealed high perceived quality (CEQ score of 12,050) and strong subjective ratings regarding students' ability to analyze, design, apply, and evaluate XAI outcomes. All students successfully completed the course, with 89% of them demonstrating confidence in multi-perspective model analysis.

Discussion: The survey results demonstrated that interactive tutorials and practical workbooks were crucial for translating XAI theory into practical skills. Students particularly valued the balance between theoretical concepts and hands-on implementation, though evaluation of XAI outputs remained the most challenging aspect, suggesting future courses should include more structured interpretation exercises and analysis templates.

KEYWORDS

XAI, ML, AI, pedagogy, education

1 Introduction

As technological advancements have increased computational hardware, modern research has resulted in high performing ML models that have found numerous applications. However, in many of these applications models are treated as black boxes, where the output in no way indicates the decision-making process behind it. Consequently, model understanding poses a notable challenge that is imperative to overcome if it is to meet the objective of responsible and beneficial use of ML systems.

To this end, the field of explainable AI (XAI) has emerged, aiming at designing tools and methodologies that allow the extraction of meaningful information out of black-box models (Arrieta et al., 2019). Although XAI is a relatively young field, it has already generated an impressive amount of scientific literature. Furthermore, there are a number of high-performance, open-source implementations of some of the most popular XAI techniques, which has facilitated their rapid adoption in commercial settings. This can also be seen in the spike of scientific publications discussing the deployment of XAI in healthcare, banking, e-commerce, cybersecurity, etc.

However, when it comes to actually employing XAI in practical applications, there is alarming evidence that professionals/data scientists use these tools in the wrong way (Kaur et al., 2020), where misuse most often arises due to misunderstandings around the scope and kind of insights that can be gained when using certain XAI techniques. Consequently, this leads to sub-optimal use of XAI, and to a misinterpretation of the resulting explanations. This finding is related to a broader issue regarding the proper use of AI/ML, which is commonly referred to as *user competence*. Despite the need for developing the technical skills required to competently use the tools provided by an AI-driven society, XAI related academic resources are extremely limited. Although there are a number of tutorial and introductory articles that can be found online, there is only a single formal academic course on XAI (Lakkaraju and Lage, 2019).

In this work we will attempt to alleviate this situation, providing a pedagogical perspective on how to structure a course on XAI, in a way that introduces students and professionals to various explainability techniques, while also keeping an eye on the big picture of the field. This proposal takes a distinct stance from the course in (Lakkaraju and Lage, 2019), since the latter focuses on introducing several technical approaches, while the presented course emphasizes the conceptual aspects of explanations, and the discussed techniques are introduced as specific realizations of broader conceptual categories. This decouples individual XAI techniques from the overall objectives, advantages, and challenges of XAI, allowing for updating the material in accordance with the new developments in the field. In particular:

- We provide a pedagogical perspective on how to structure a course in XAI.
- We pair each lecture with a series of open-ended questions to promote an exchange of ideas between lecturers and/or participants.
- We develop a series of technical tutorials that discuss practical implementations of XAI techniques.
- We evaluate the course based on the feedback and assignments provided by the MSc students that attended the course, as it was offered by the University of Edinburgh.

2 Materials and methods

2.1 Learning objectives

Probably the most common way to introduce XAI to individuals interested in applying related techniques is through tutorials, for example Bennetot et al. (2021); Rothman (2020). However, most of them are targeted toward a technical research audience, where the purpose is research discovery among peers and not the teaching of fundamentals. Keeping in mind that tutorials usually serve as short, technical introductions to a subject, providing detailed insights regarding the nuances of different techniques or explanations styles is beyond their scope.

One of the main goals of the course is to fill this need by imparting a number of key concepts. The first one is that the various explainability approaches can be taxonomised such that a technique can be selected by considering explanation types, explanation properties, advantages and disadvantages, as well as the specific model under consideration. Another important concept is that XAI can be used to structure a narrative in order to successfully answer potential questions raised by stakeholders, in line with evidence that explainability techniques are best linked to stakeholder questions (Arya et al., 2019). Along with imparting theory, an equally important goal is to provide students with experience in applying these explainability techniques using commonly available APIs, so they can gain an understanding of the implementation pipeline (e.g., data cleaning, parameter tuning, etc.).

In broader terms, this course enables inclusivity, empowerment, and responsibility with respect to XAI. In regards to *inclusivity*, the course suggests strategies that can help make XAI more easily understood/accessible, especially since current developments can be very technical and difficult to grasp for those not keeping up with the state of the art. Focusing on a representative subset of techniques and showing that they can be tightly coupled with certain types of questions, provides an accessible strategy for introducing students/practitioners into the field. With respect to *empowerment*, the course is designed for students with some experience with data; however, it also includes preparatory lectures on machine learning. It gives students the opportunity to engage with machine learning models, debug them, and inspect whether the resulting models fit their purpose. In terms of enabling *responsibility*, it is widely acknowledged that responsible design in artificial intelligence includes many facets, from bias detection to value alignment. In this broad picture of ensuring that machine learning models perform as they should, explainability is an essential ingredient, so it is important that practitioners can use such tools competently.

To the best of our knowledge, this is the first work on this topic, so it is not possible to make empirical comparisons. However, the course was evaluated based on students' final assignment, as well as their feedback, achieving very positive results. Hopefully, this will be the start of a discussion on how to effectively teach this very important topic.

Based on our experience developing and delivering this XAI course, we propose a three-pillar pedagogical framework that can be adapted for similar courses. The framework allocates course content as follows: Conceptual Foundation (30% of course time) establishes the theoretical groundwork by distinguishing between transparent and opaque models, introducing XAI taxonomy, and explaining different transparency levels before examining specific techniques; Technique-Specific Modules (50% of course time) provide in-depth coverage of 3–4 core XAI techniques (such as SHAP, Counterfactuals, and InTrees) complemented by 2–3 additional techniques (such as Anchors, PDP/ICE, and Deletion

Diagnostics) that demonstrate different explanatory perspectives; and Practical Integration (20% of course time) through hands-on tutorials and a comprehensive final project requiring students to apply multiple techniques to justify model acceptance or rejection.

Upon completion of this course, students are expected to have learnt to apply XAI techniques to examine a given model, as well as to have gained an overview understanding of the conceptual distinctions of explanations. More specifically, the expected learning outcomes are as follows:

- **Analyse:** Describe the context of the machine learning application and why explainability would help, but also scrutinize which kind of explainability technique is necessary.
- **Design:** Define the implementation pipeline for an application; provide a means to clean the data, install and set up one or more *posthoc* explainability techniques.
- **Apply:** Competently apply a wide range of techniques and tools, also knowing their particular features and drawbacks. Have the foundations to understand new and upcoming methods and techniques.
- **Evaluate:** Critically reflect on the results of XAI techniques and investigate their utility in the given context.

In particular, since both the theoretical and practical aspects are covered, students should be able to understand the context in which these techniques are deployed but also understand the theory in justifying these techniques. The final project in particular is an opportunity for students to create a narrative and an application and motivate and argue for or against a model by using a sequence of techniques.

2.2 Course structure and content

The course consists of 9 lectures paired with 4 practical tutorial sessions, delivered over one semester. The lectures cover the following topics: (i) ML preface, (ii) XAI preface, (iii) SHAP, (iv) PDP/ICE, (v) Counterfactuals, (vi) Anchors, (vii) Deletion diagnostics, (viii) InTrees, and (ix) Future research directions. Each major XAI technique receives both theoretical treatment in lectures and practical implementation through dedicated computational notebooks, where students engage in coding exercises and raise practical issues about the corresponding techniques. Students are expected to submit 2 assignments and a final project. The course culminates with an open-ended final project where students must select their own dataset and model, then employ multiple XAI techniques to construct arguments for stakeholder decision-making. This structure ensures students develop both conceptual understanding and practical competency while learning to integrate multiple explanation approaches.

3 Course content

In what follows, we outline the structure of the course, which is comprised of 9 lectures as well as 4 tutorials over the semester, while the students are expected to submit 2 assignments and a final project. Each lecture is focused on a specific topic, such as detailing the characteristics, advantages, and disadvantages of a certain XAI technique. Each lecture concludes with three carefully designed open-ended questions that promote discussion and critical thinking around the topic at hand. Our question framework follows three categories: Comparative Questions that encourage students to analyse trade-offs between different approaches (e.g., "What are the trade-offs between TreeSHAP and KernelSHAP?"); Application Questions that challenge students to extend techniques to new contexts (e.g., "How would counterfactuals work with image data?"); and Critical Thinking Questions that probe assumptions and potential failure modes (e.g., "How could a biased trained model 'trick' SHAP by hiding its bias?"). These questions are designed without definitive "correct" answers, encouraging students to engage in meaningful discussion and develop nuanced understanding. Questions consistently connect specific techniques to broader issues in AI ethics, fairness, and responsible deployment.

3.1 Preliminaries

3.1.1 ML preface

We begin by briefly introducing ML objectives, focusing on the various stages of the general process (i.e., acquiring a dataset, selecting and training a model, and evaluating its performance). Following that, we discuss the properties of certain ML models that are going to be used as a reference in the next lectures, in order to demonstrate XAI concepts. Specifically, we make a distinction between transparent models, such as Linear/Logistic regression, and opaque models, such as Neural networks.

The goal of this lecture is to highlight which properties make a model more understandable to a human, emphasizing intuition. This serves as a first step toward understanding the necessity of developing XAI techniques in order to enhance opaque models with interpretable features, so they resemble their transparent counterparts.

Open ended questions, such as those below, are presented to students for debate and discussion; such questions are discussed in each lecture: *(a) Can you think of a reason why opaque models, often, have better performance than transparent models? (b) Can you think of cases when opaque models are outperformed by transparent models? (Hint: consider relational data.) (c) Can you argue about why we may want to solve an ML challenge using both transparent and opaque modeling?*

3.1.2 XAI preface

In this lecture, we begin with discussing cases that highlight the need for XAI, such as adversarial examples that drastically alter the model's original outcome, while being indistinguishable from humans. Following that, we talk about the various transparency levels that a model can satisfy (Arrieta et al., 2019): **Simulatability:** A model's ability to be simulated by a human. **Decomposability:** The ability to break down a model into parts (input, parameters, and computations) and then explain these parts. **Algorithmic transparency:** The ability to understand the procedure the model goes through in order to generate its output.

We emphasize the different requirements of each level, as well as cases where a model could satisfy all of them (such as simple linear models), or none of them (such as multi-layer neural networks). This leads to the observation that, in the latter case, *post − hoc* methods for inspecting the internal mechanisms of a model are essential to employ them in critical applications. In general, XAI is important in identifying issues such as: **User acceptance:** By providing explanations, users are more likely to be satisfied and accept an ML decision. **Improving human insight:** Beyond just using ML to perform automation tasks, scientists can use ML for research purposes with respect to big data. An intelligible model can provide information to scientists based on the data being modeled. **Legal imperatives:** Using ML to assess legal liability is a growing issue, as auditing situations to determine liability requires clear explanations from a model's decision. The European Union's GDPR legislation decrees citizens' right to an explanation further strengthening the need for intelligible models.

After the importance of XAI is established, the next step is to provide an overview of the ways that have been considered for producing explanations. The first distinction is between global explanations, that explain the model as a whole, and local ones, that attempt to explain the model's prediction for the specific datapoint of interest. In addition, another distinction comes based on the applicability of each technique, where model-agnostic ones are generally applicable to any model, while model-specific techniques are designed to be applied in certain classes of models. Finally, we give examples of how feature relevance scores, visualizations, examples, and text can be used to form explanations (Arrieta et al., 2019).

Open ended questions to present at the end of the lecture include: *(a) Can you argue how the XAI evaluation criteria differ or not from the criteria for a "good" ML model? (b) If a medical system offers* 98% *accuracy over a transparent model that only offers* 88% *accuracy, what might you prefer, and why? (c) Would an ensemble of different transparent models be considered transparent?*

3.2 XAI techniques

Here we propose an indicative set of XAI techniques that could be taught in the course. We begin with three techniques we consider to be essential (SHAP, counterfactuals, and InTrees), each of them for its own reasons. SHAP is arguably the most popular XAI approach, having well founded theoretical properties, so we believe it is important for students to be familiar with it. On the other hand, counterfactuals bring together philosophy and XAI, approaching the problem from a completely different angle, while also serving as a stepping stone to more advanced causal inference concepts. Finally, InTrees is a model-specific technique that clearly demonstrates the advantages of utilizing them over model-agnostic ones, while it is also simple and relatively straightforward.

In addition to the above, we briefly discuss other techniques that could be taught along the three main ones. Their selection was determined by our narrative and how we believe an introduction to XAI should be approached, based on Belle and Papantonis (2020). This is why we complement the aforementioned techniques with Anchors, which produce simple propositional rules, and visualizations, which are especially valuable when communicating explanations to non-technical audiences, as well as deletion diagnostics, which considers the model as a function of the training dataset. Each of these techniques brings a different perspective on XAI, which we think is necessary to form a complete picture of the underlying model's reasoning. A different narrative might motivate more global techniques, deep learning techniques, or symbolic approaches. Moreover, even with the same narrative, we could replace certain techniques with similar ones, for example, anchors with LIME. This means there is a lot of flexibility, depending on the students, the goals of the course, and/or the preferences of the instructor. Having said that, Anchors, visualizations, and deletion diagnostics are the techniques we included when delivering the course, so we will present some of the details of the context and style of the lectures since they probably played a role in the students' answers to the questionnaire they completed at the end of the course.

3.2.1 SHAP

Shapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017) is a model-agnostic method for explaining individual predictions. SHAP learns local explanations by utilizing Shapley Values (Shapley, 1952) from co-operative game theory, in order to measure feature attributions. The objective is to build a linear model around the instance to be explained and then interpret the coefficients as feature importance scores.

Shapley values provide a means to attribute rewards to agents conditioned on the agent's total contribution to the final reward. In a cooperative setting, agents collaborate in a coalition and are rewarded with respect to their individual contributions. In order to apply this technique to ML models, it is necessary to make adjustments so the problem is expressed in a game theoretic manner: **Setting/Game**: SHAP interprets the model prediction on a single input, x, as a game. **Agents/Players**: The different features of input x are interpreted as being individual players. **Reward/Gain**: Measured by taking the model prediction on the input x and subtracting the marginal predictions, i.e. predictions where some of the features are absent.

Having made these adjustments, the Shapley value of feature i equals:

$$\phi_i = \sum_{S \subset F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)]$$

where F is the set of all features, $f_{S \cup \{i\}}(x_{S \cup \{i\}})$ is the model's decision when the features in $S \cup \{i\}$ are given as input, and $f_S(x_S)$ is the decision when only features in S are given.

From a pedagogical perspective, introducing SHAP comes with a number of benefits, such as: (i) It exemplifies how well established mathematical ideas can be adjusted to take on new problems, demonstrating the multidisciplinary nature of ML related research. (ii) Shapley values are known to satisfy some important properties, allowing for a discussion focused on why these properties are important when generating explanations, or why it is important to have such theoretical guarantees. (iii) The current implementation of the SHAP python module comes with an array of different

visualizations, which the students can inspect in order to strengthen their understanding of SHAP.

Open ended questions include: *(a) What are the trade offs and differences between TreeSHAP and KernelSHAP? (b) How could a biased trained model "trick" SHAP by hiding its bias, i.e. assign Shapley values to protected features that do not match their actual importance in the model's decision? (c) With KernelSHAP, the sampling for missing values assumes feature independence, is there a way to remedy this issue? Can you think of possible solutions?*

3.2.2 Counterfactual

A counterfactual explanation is a statement that identifies how a given prediction would need to change for an alternate outcome to occur. Key to counterfactuals is the idea of "the closest possible world" which signifies the smallest possible change on a set of variables that suffices to alter a model's outcome (Lewis, 1973). For example, if a loan application is rejected, then providing a counterfactual (i.e., a successful application which is as similar as possible to the original one), makes it easier for a person to identify the important information that is relevant to their specific application. In a sense, counterfactuals highlight why a decision was not made, in contrast to other approaches that aim at explaining why a decision was made.

One of the most popular frameworks for generating counterfactuals for ML models is based on Wachter et al. (2018), where the authors express the problem as:

$$\min_{x} d(x, x_i) \text{ s.t.}$$
$$f(x) = Y$$

where d is a distance function, x_i is the factual datapoint, x is the counterfactual one, $f(\cdot)$ is the ML model, and Y is the category we would like the counterfactual point to be classified into. For differentiable models, this problem can be solved using Lagrange multipliers, along with an optimisation scheme, such as ADAM (Kingma and Ba, 2015).

Introducing counterfactuals is beneficial from a pedagogical perspective since (i) They provide an entry point for drawing connections with concepts from causal inference (CI). CI is expected to be one of the most promising future research directions, but it is often challenging for students to grasp the underlying concepts. However, the notion of counterfactuals used in XAI is a simplified version of the ones in CI, so it is possible to build on them in order to facilitate the understanding of more advanced ideas. (ii) In addition, discussing the progression from the initial work in Wachter et al. (2018) to more recent advances demonstrates how XAI is a dynamic field, where a technique can be refined by taking into account new requirements or desiderata. This could help develop students' critical thinking, enabling them to identify the reasons why such progressions happen. (iii) Finally, counterfactuals showcase the interplay between XAI and other domains, such as fairness in AI, or applications, like model debugging, all of which exemplify the interdisciplinarity of XAI related research. For example, by probing a model by generating multiple counterfactuals, we can examine whether changes in sensitive attributes (such as gender) may lead to the model producing a different outcome. If this is the case, then this is a clear indication that the model exhibits biased behavior.

Open ended questions include: *(a) Can you get multiple counterfactuals for a given instance? If yes, how should we interpret them? (b) How can we handle discrete features? (c) How would counterfactuals work with image data?*

3.2.3 InTrees

InTrees (Deng, 2014) are a model-specific XAI method for tree ensembles, which take advantage of the tree architecture to produce interpretable "if-then" explanations. One of the advantages of this technique is that it is based on intuitive and easy to explain algorithms, although we are not going to get into the details in this work. In turn, InTrees demonstrate how certain black-box architectures may contain pieces of information that can facilitate the model's understanding. It is worth noting that at the core of this technique lies the idea that although tree ensembles might be opaque, each of its constituents is transparent, so they can be readily inspected.

The previous observation perfectly captures the utility of model-specific techniques; instead of relying on universal approximations, develop alternatives that take advantage of the specific characteristics of the model at hand. The majority of model-agnostic approaches make significant assumptions about the underlying model, which are often violated, compromising the quality of the resulting explanations. Consequently, one of the main drives of model-specific explanations is to reduce the number of assumptions, leading to more accurate explanations. Introducing InTrees has the benefit of clearly demonstrating the concept in a simple way, as opposed to more mathematically challenging alternatives, for example, neural network LRP explanations (Bach et al., 2015). This should improve the student's understanding of why model-specific explanations are important, without getting into overly complex technical details.

Open ended questions include: *(a) What is the tradeoff between frequency and error in practical scenarios? Which should we aim to optimize? (b) Can you argue with examples about what happens if pruning is not applied? (c) Can a similar idea be applied to non-tree ensembles (e.g., SVM, neural networks)? If so, how do you think this would be possible?*

3.2.4 PDP/ICE

Another prominent means of explaining an ML model is using visualizations, especially when communicating explanations to a non-technical audience. A Partial Dependence Plot (PDP/PD) (Friedman, 2001) plots the average prediction for a feature(s) of interest as the feature's value changes. These plots can reveal the nature of the relationship between the feature and the output, for example, whether it is linear or exponential. PDPs present global explanations, as the method factors all instances and provides an explanation regarding the (marginal) global relationship between a feature and the model prediction. Assuming we are interested in examining the partial dependence of the model f on a feature s, we

have to compute:

$$f^*(x_s) = \sum_{i=1}^{N} f(x_s, \mathbf{X}_{-s}^{(i)})$$

where N is the cardinality of the dataset, and $\mathbf{X}_{-s}^{(i)}$ is the $i-th$ datapoint, excluding feature s. We see that the partial dependence function provides the average marginal output for a given value of x_s. Furthermore, it is not difficult to extend this method to account for the partial dependence of a function on more than one feature, however, this is usually done for one or two features, due to our inability to perceive more than 3 spatial dimensions.

Individual Conditional Expectation (ICE) (Goldstein et al., 2013) plots model predictions for multiple instances, where for all instances, only the feature of interest changes value, while the remaining feature values for a given instance are held constant. This plot shows the feature-value and model prediction relationship, for each instance as a separate line, as opposed to the single average predictive line with PDP. Some of the advantages of employing ICE are: (i) PDPs may hide some of the heterogeneous relationships between feature value interactions by averaging them out. Consider that two opposing, but equally valued, influences on model prediction can be canceled out when averaged. (ii) PDPs provide a view into the average predictive behavior of a model with respect to a single feature, but the validity of this predictive behavior is diminished by any possible interactions with the feature being plotted and the remaining features in the data. (iii) ICE plots are able to plot more accurate relationships even with the presence of highly correlated features.

Finally, there is an interesting relationship between these two plots, as averaging the ICE plots of each instance of a dataset, yields the corresponding PD plot.

Open ended questions include: *(a) What are the key limitations of PDP/ICE? (b) Roughly sketch a 3-dimensional PDP and ICE plot. Based on that, argue about whether this makes explaining the model easier or not. (c) Instead of averaged out values, how can you show the minimum and maximum for the features in PDP? Is that useful?*

3.2.5 Anchors

Rule-based classifiers have been traditionally utilized due to their transparency since they are easy to inspect and understand. Anchors (Ribeiro et al., 2018) is an XAI technique that builds on this principle, aiming at generating simple rules to describe a model's reasoning. They explain individual predictions locally by identifying a decision rule that "anchors" the prediction in question, thus they operate on an instance level. A rule anchoring a prediction implies that changes to the remaining feature values do not impact the prediction. A rule's *coverage* is defined as the fraction on instances that satisfy the "if" part of the rule. Moreover, a rule's *precision* is the fraction of instances that satisfy both the "if" and the "then" part.

Formally, an anchor, A, is defined as the solution to the following problem:

$$\max_{\text{A s.t. } P(prec(A)\geq\tau)\geq 1-\delta} cov(A)$$

where $prec(A) = \mathbb{E}_{D(z|A)}(\mathbf{1}_{f(x)=f(z)})$ is the precision, $D(z|A)$ is the data distribution given the anchor, $f(z)$ is the ML model, $\mathbf{1}$ is the indicator function, and $cov(A) = \mathbb{E}_{D(z)}(A(z))$ is the coverage. In words, this optimization problem looks for if-then rules where the preconditions (the "if" part) contain conditions that are satisfied by as many instances as possible, while requiring that these points also satisfy the "then" part, with high probability. This way the resulting rules are not based on niche characteristics of the specific datapoint at hand but are as generally applicable as possible.

Open ended questions include: *(a) What would you choose between anchors with high precision and low coverage vs anchors with low precision and high coverage? (b) Can you give examples of rules that might apply to recent data you have encountered? Can you argue about what precision/coverage you expect them to have? (c) Compare anchors to other local explainability techniques, such as SHAP. What are the advantages and disadvantages compared to it?*

3.2.6 Deletion diagnostics

Deletion diagnostics is a technique which investigates the model as a function of its training data. It considers the impact of removing a particular training instance from the dataset on the final model (Cook, 1977). By removing an instance with significant influence from training, deletion diagnostics can help with model understanding. In this context, an instance is considered to be influential if its removal causes the parameters of the trained model to change significantly or results in notably different predictions on the remaining instances.

Identifying influential instances is important since they invert the relationship between model and data where we now look at the model as a function of the training set. Influential instances can inform how specific feature values influence model behavior, they can also be used to identify adversarial attacks, they can help in debugging by identifying instances which result in model errors, and they can help in fixing mislabelled data (Koh and Liang, 2017). All of these are significant information when explaining a model, which could also find additional applications, for example reducing the size of the training dataset. This could be achieved by retaining only the influential instances and then retraining the model using them, since these instances can express model behavior where it is most sensitive, contributing to a better understanding of model behavior.

Influence functions (Koh and Liang, 2017) are a contemporary alternative to standard methods of deletion diagnostics, wherein the removal of an instance i is approximated, and the model does not need to be retrained with instance i removed. This makes it more efficient to estimate the influence of a datapoint on the final model, since, retraining a model every time an instance is removed is very computationally intensive.

Open ended questions include: *(a) Can you explain how deletion diagnostics can be done efficiently without retraining? (b) Do you think deletion diagnostics can be applied to random forests? (c) What do you think is the relative usefulness of deletion diagnostics compared to influence functions?*

3.2.7 Future research directions

The final lecture of the course is about the future of XAI related research. Its goal is to discuss the limitations of XAI and

prepare students for the next generation of techniques, as well as to provide an overview of which concepts are likely to play a central role in the future. This way the interested students have the chance to study these concepts in advance, so they have the prior knowledge required to potentially grasp future techniques. Of course, it is not possible to be exhaustive and cover all directions, instead, we provide an indicative list of current XAI limitations.

An important point of emphasis for students is to realize that limitations arise both on a technical and a practical level. Most of the existing techniques, especially model-agnostic ones, require resorting to approximations. This means that there is always the danger that the resulting explanations might not be accurate, or even be misleading. Furthermore, existing approaches are not really able to identify spurious correlations and report them back. Due to this, it is possible for features to look like they have a strong influence on each other, when, in reality, they only correlate due to a confounder. A possible resolution to these issues could be introducing more concepts from causal analysis, which is already a major drive in related areas, such as fairness in ML. For example, if an explanation was accompanied by a causal model it would not be difficult to check for any spurious correlations.

On a more practical level, developing XAI pipelines to explain a model it is still an open research question. Currently, there is no consensus regarding either the characteristics of a good explanation or the way of combining existing techniques in order to adequately explain a model. While there is some overlap between the various explanation types, for the most part, they appear to be segmented, each one addressing a different question. This hinders the development of pipelines that aim at automating explanations or even reaching an agreement on how a complete explanation should look like. On top of that, it is not clear whether explanations should be *selective* (focus on primary causes of the decision making process), or *contrastive* (indicate why a model made decision X, and provide justification for deciding X rather than Y), or both, and how to extract such information from current techniques. Audiences in XAI can include experts in the field, policy-makers, or end users with little ML background, so intelligibility should be varied in its explanations depending on the knowledge level and objectives of the audience. Interdisciplinary research combining psychology, sociology, and cognitive sciences can help XAI in delivering appropriate explanations (Miller, 2019).

Open ended questions include: *(a) What do you think are the most important limitations of XAI? (b) Can you suggest ways to automate XAI? (c) Can you suggest ways to address the potential dangers of transparency?*

3.3 Assignments

While the course content includes general information as well as mathematical theory, students' assessment is purely practical. As the course is aimed at industry practitioners, a greater emphasis is given to applying the XAI techniques in a real world setting. Variants of the course focused on teaching MSc students might emphasize the various algorithmic formula and mathematical derivations for purposes of developing future XAI researchers.

Effective XAI pedagogy requires careful integration of theoretical concepts with immediate practical application. We recommend pairing each theoretical concept with hands-on implementation exercises, for instance, following the mathematical presentation of SHAP with TreeSHAP notebook exercises using real datasets. The use of realistic, ethically-relevant datasets naturally connects technical implementation to broader discussions of responsible AI and algorithmic fairness. Our assignment structure maintains a 60% technical implementation and 40% interpretation balance, ensuring students not only learn to execute XAI techniques but also develop the critical thinking skills necessary to interpret and communicate results.

The assignment structure includes four jupyter notebooks (Kluyver et al., 2016) namely SHAP, counterfactuals, Anchors, and InTrees (PDP/ICE and deletion diagnostics omitted for simplicity), each of which demonstrate a singular XAI technique, which provides questions that students must analyse and answer for assessment. Questions include technical code based implementations of a specific XAI technique and short answer questions asking for student interpretation of the outputs. A final project is also assigned which asks students to select a dataset and model of their choice and describe a problem requiring explainability, and consequently implement a minimum of two *post−hoc* explainability methods. The application of these methods will be used as evidence by students in their discussion of the model's performance on the dataset. Here, the results and the implications from the various XAI techniques will be used by the students to convince a stakeholder to either accept or reject a model.

3.3.1 Workbooks

At the beginning of each tutorial, students need to import the relevant python libraries (`pandas`, `numpy`, etc.) as well as the library associated with the corresponding XAI technique. Following the initial setup, basic ML preprocessing practices such as handling *missing values, visualizations,* and *feature engineering*, are applied.

Our computational notebooks follow a consistent four-section template: Setup Section covering library imports and data preprocessing to build general ML pipeline skills; Demonstration Section providing guided implementation with detailed explanations; Assignment Questions mixing technical implementation tasks (3–4 questions) with interpretation exercises (2–3 questions); and Extension Challenges requiring students to apply the technique to different models or datasets.

The next step is *Modeling*, which includes training the model and assessing its performance. Each tutorial uses an arbitrary model, the majority of which are black-box models, and performance is measured by looking at the accuracy, precision, recall, and f1 scores.

The third section of the tutorial is where the corresponding XAI technique is applied. The SHAP and Counterfactual tutorials end with a series of assignment questions which comprise a portion of the submission work. Assignment questions include 2–3 short answer questions regarding the technique and/or potential concerns in applying a given model, followed by 3–4 technical questions where the student will need to apply the XAI technique either on another model or on an augmented dataset.

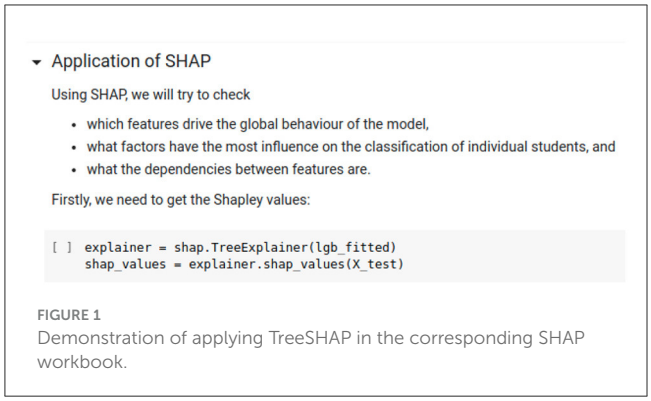

FIGURE 1
Demonstration of applying TreeSHAP in the corresponding SHAP workbook.

Assignment questions are carefully categorized by type: Technical Questions require implementation skills (e.g., "Apply XGBoost with SHAP to the Portuguese scores dataset and identify the 5 most important features"); Interpretation Questions develop communication abilities (e.g., "Write a paragraph for a non-technical audience explaining how your model makes decisions based on SHAP outputs"); and Comparative Questions build analytical thinking (e.g., "Compare SHAP importance rankings across different model types and discuss any discrepancies"). This structure ensures students develop both technical competency and the interpretive skills necessary for real-world XAI applications. In the remainder of this section, aspects of the SHAP tutorial are presented to give a representative example of the structure/content of each tutorial.

SHAP: The first step is to import the relevant library (`!pip install shap`) and, after preprocessing the data and training a LGBM model, SHAP is invoked. As LGBM is a tree ensemble method, the model-specific SHAP implementation is utilized (see Figure 1).

Included throughout the tutorial are summary questions, which are designed as general discussion points following the introduction of an ML or XAI technique. Questions $(1-2)$ are specific to assessing the black-box model, while questions $(3-10)$ are specific to SHAP and its implementation. In general, these questions are discussed with students during tutorial sessions.

1. How would you interpret 50% precision in the table above?
2. Which metric from the above do you think would be of the most interest to a stakeholder interested in a model that aims to predict students' performance?
3. Generate the same plots for instances in class 0. What is the link between these graphs and what has been generated for class 1?
4. What are the 5 most important features which are driving this model's decisions?
5. How would you interpret the horizontal axis of the first summary plot above? What does a SHAP value of -1.5 mean?
6. Can we express SHAP values in terms of probability? Justify your answer.
7. According to these plots, which student is most likely to fail (assuming our model is appropriate)?
8. Are there any features which are important from the perspective of predictions on the local level for these 3 students and which could indicate some fairness issues?
9. Produce a local plot as above for another student (feel free to select your own observation). How different is it from the global picture?
10. What does the horizontal spectrum on the top of the graph show? What do those values mean?

At the end of the SHAP workbook, we provide students assignment questions which are to be submitted for assessment. These questions will require the students to implement the techniques previously demonstrated in the workbook as well as require investigations online for methods not shown. Overall, questions can be separated into technical (coding focused) and short answer. For the SHAP workbook we classify questions (1–5) as being primarily technical and questions (6–8) as short answer, requiring student interpretation.

1. Use the public dataset introduced in this tutorial and apply an XGBoost model. Your outcome variable will be Portuguese language scores pooled into class 0 and 1 in the same way as in this notebook (feel free to skip any hyperparameter tuning operations). Make predictions on your test set and produce a set of measures that describe the model's performance.
2. Using SHAP summary plots, what are the 5 most important features in the model?
3. Create a decision plot for all observations and all features in your test set, highlight misclassified observations and create decision plots for the set of misclassified observations and for 4 single misclassified observations. Then include force plots for all observations as well as for the set of misclassified observations.
4. Make SHAP dependence plots of the 4 most important features. Use sex as a feature possibly influencing SHAP outputs. This is done by setting the `interaction_index` as "sex."
5. In the light of the plots from 3 and 4, discuss whether the interaction effect between sex and other features can meaningfully impact decisions of your model.
6. Discuss how various SHAP-based graphs can be used in the process of model validation.
7. Write a paragraph for a non-technical audience explaining how your model makes decisions based on SHAP outputs. Ensure the text is clear of jargon!

3.3.2 Final project

The final project requires students to consider an ML application, and then carry out all the necessary steps to train a model. After the training is completed, students need to utilize a series of XAI techniques to evaluate the resulting model and argue about whether it should be retained or dismissed. The minimum time commitment expected from students is 14 hours. Essentially all aspects of the problem specification are decided by the students, i.e. dataset selection, model selection, and XAI techniques.

To improve stakeholder communication skills, we recommend adding argument construction modules to each major technique tutorial. These modules should include: (1) Stakeholder Persona Templates that outline different audience needs (e.g., technical teams want precision metrics, executives want business impact, regulators want fairness evidence); (2) Translation Examples showing how to convert technical XAI outputs into

stakeholder-relevant insights (e.g., "SHAP analysis reveals that credit history accounts for 40% of loan decisions, suggesting our model aligns with traditional banking practices"); and (3) Argument Structure Frameworks providing templates like "Problem → XAI Evidence → Business Implication → Recommendation."

Furthermore, students need to come up with a narrative describing the problem, for example, *"Taking a credit scoring dataset, and the XGBoost model, convince a banking institution to reject the model using (at least) technique 1 and technique 2"*. The goal of the project is to prompt students to use various XAI techniques in order to convince a (hypothetical) stakeholder to approve/dismiss the underlying model. This situation resembles what students might come across when applying XAI in a professional setting, so it is important that they can form sound arguments based on the explanations at hand.

To build application confidence, we recommend implementing progressive scaffolding exercises that gradually transition from guided tutorials to independent application. This includes: (1) Bridge Assignments where students apply tutorial techniques to slightly modified datasets or models, reducing the cognitive load of simultaneous technique learning and independent application; (2) Troubleshooting Workshops that explicitly address common implementation challenges (e.g., handling categorical features in SHAP, interpreting counterfactuals for imbalanced datasets) with step-by-step debugging approaches; and (3) Technique Selection Guides that provide decision trees helping students choose appropriate XAI methods based on their model type, data characteristics, and stakeholder questions. Additionally, the course provides ongoing feedback through assignment submissions and tutorial discussions, where students receive guidance on technique selection and implementation strategies, building confidence through collaborative problem-solving during tutorials.

3.4 Evaluation methodology

To assess the effectiveness of the course, a modified version of the Course Experience Questionnaire (CEQ) and a general performance analysis based on the coding assignment were used.

3.4.1 Context and participants

Sixteen survey responses were collected from the students who attended the presented course as part of their MSc studies at the University of Edinburgh. Eight of the respondents were postgraduate research students, five were doing non-academic work, while three responded "Other." Prior to taking this course, 87.5% of respondents were not familiar with the topic of XAI. Most of the respondents had at least some practical experience of working with ML (81.3%) and theoretical knowledge of ML (81.3%). One respondent had no theoretical or practical knowledge of ML, and two respondents had some theoretical, but no practical ML experience. Four respondents rated their practical experiences in ML as close to expert level, and six rated their theoretical ML knowledge close to expert level.

3.4.2 Procedures and data collection

Conventionally the CEQ is employed as a measure of teaching quality of university courses (Ramsden, 1991). The questionnaire assumes a strong connection between the quality of student learning and student perceptions of teaching (McInnis, 1997). When combined with the additional course assessment items and adapted to the course context it can reliably be applied as a domain-neutral indicator of university course quality (Griffin et al., 2003). For this study, 12 items were selected from the original version of the CEQ based on their relevance to the XAI course context (see the Supplementary material). These items were scored on a 5-point Likert-type rating scale from "strongly disagree" to "strongly agree."

In addition to the CEQ items, context-specific items were used in order to gather information about (i) pre-existing skills and theoretical knowledge of XAI; (ii) pre-existing skills and theoretical knowledge of Machine Learning (ML); (iii) satisfaction of the diversity of the taught techniques; (iv) course ability to build an understanding of XAI techniques; (v) level of comprehension of the conceptual distinctions, advantages, and disadvantages of the XAI techniques covered in the course; (vi) success in meeting the four pre-set learning objectives. These responses were also based on a five-point Likert scale. Finally, in an open-ended manner, students were asked to list their favorite aspect of the course and to suggest anything that could help to improve the course in future. The resulting questionnaire was given to the students after concluding the final lecture, but before the results of the final assignment came out, so they were not aware of whether they had passed the course.

4 Results

The results were divided into two parts. Overall CEQ scoring of the responses, as well as quantitative and qualitative analysis of the individual statement ratings. The CEQ raw scores were recorded as follows: a raw score of 1 ("strongly disagree") was recoded to −100, 2 to −50, 3 to 0, 4 to 50, and 5 ("strongly agree") to 100, eliminating the need for decimal points. The scoring of negatively worded items was reversed. In interpreting CEQ results, a negative value corresponds to disagreement with the questionnaire item and a positive value to agreement with the item. Positive high scores indicate high course quality as perceived by graduates. The responses revealed a high positive overall CEQ score of 12, 050, which indicated high course quality as perceived by respondents. The CEQ is widely accepted as a reliable, verifiable, and useful measure of the perceived course quality (Griffin et al., 2003).

On top of that, in order to gain a more fine grained picture of students' responses, individual questions 4, 5, 6, 7, 16, 17, 19, 20, 21, 22, 23 were analyzed. Questions 20 − 23 correspond to the learning objectives (Analyse, Design, Evaluate, Apply), while the remaining ones correspond to items indicating the level of students' satisfaction/confidence (see Table 1). The analysis was performed by estimating a 95% confidence interval for the average score of each of these questions. Each interval was constructed using the non-parametric bootstrap method (Efron and Tibshirani, 1986). Figures 2, 3 show the obtained results,

Starting from Figure 2, questions 16, 17 examined whether the course was overly theoretical or practical (respectively), so the fact that they have both received a score of about 3, indicated that

TABLE 1 A description of the questions included in the analysis.

Questions	1	2	3	4	5	Figure Ref.
Student confidence and understanding						
6. Please rate how confident do you feel in applying the XAI techniques you learned in your own models	**Not at all** – I do not feel I can apply **any** of the techniques I learned to my models		**Somewhat** – I feel I can apply **most** of the techniques I learned to my models		**Very** – I feel I can apply **all** the techniques I learned to my models	Figure 2
7. Please rate how satisfied are you from the diversity of XAI techniques covered in the course	**Not at all** – The techniques were **overly similar**		**Somewhat** – The techniques were somewhat diverse, but there was **significant overlap**		**Very** – The techniques were very diverse, and had **minimal overlap**	Figure 2
8. Please rate how much do you feel your understanding of XAI was benefited by the course	**Not at all** – I do not feel the course helped me understand XAI **at all**		**Somewhat** – I feel I now understand **some** aspects of XAI better		**Very** – I feel I now understand **many** aspects of XAI better	Figure 2
9. Please rate how much do you feel you have comprehended the conceptual distinctions, advantages, and disadvantages of the XAI techniques covered in the course	**Not at all** – I do not feel I have comprehended **any** of these, for **any** technique		**Somewhat** – I feel I understand **some** of these, for **some** of the techniques		**Very** – I feel I understand **all** of these, for **all** the techniques	Figure 2
Course structure assessment						
16. The course was overly theoretical and abstract	**Strongly Disagree** – The course was not theoretical **at all**		**Neutral** – It had some theoretical aspects, but it was **not excessive**		**Strongly Agree** – The course was **overly theoretical**	Figure 2
17. The course was overly practical	**Strongly Disagree** – The course was not practical **at all**		**Neutral** – It had some practical aspects, but it was **not excessive**		**Strongly Agree** – The course was **overly practical**	Figure 2
19. Overall, I am satisfied with this course	**Strongly Disagree** – I am **not** satisfied at all with this course		**Neutral** – I am **somewhat** satisfied with this course		**Strongly Agree** – I am **satisfied** with this course	Figure 2
Learning objectives achievement						
20. **Analyse:** Describe the context of the machine learning application and why explainability would help, but also scrutinize which kind of explainability technique is necessary.	**Unsuccessful** – The course did **not improve** my abilities to analyse a problem		**Somewhat successful** – The course **moderately improved** my abilities to analyse a problem		**Successful** – The course **significantly improved** my abilities to analyse a problem	Figures 3, 4a
21. **Design:** Define the implementation pipeline for the project: provide a means to clean the data, install and set up one or more *posthoc* explainability techniques through a self-chosen set of programming platforms.	**Unsuccessful** – The course did **not improve** my abilities to design a pipeline for a problem		**Somewhat successful** – The course **moderately improved** my abilities to design a pipeline for a problem		**Successful** – The course **significantly improved** my abilities to design a pipeline for a problem	Figures 3, 4b
22. **Evaluate:** Critically reflect on the results from such techniques and suggest how it helps the problem context.	**Unsuccessful** – The course did **not improve** my abilities to evaluate the obtained results		**Somewhat successful** – The course **moderately improved** my abilities to evaluate the obtained results		**Successful** – The course **significantly improved** my abilities to evaluate the obtained results	Figures 3, 4c
23. **Apply:** Competently apply a wide range of techniques and tools, also knowing their particular features and drawbacks. Have the foundations to understand new and upcoming methods and techniques.	**Unsuccessful** – The course did **not improve** my abilities to apply or understand new techniques		**Somewhat successful** – The course **moderately improved** my abilities to apply or understand new techniques		**Successful** – The course **significantly improved** my abilities to apply or understand new techniques	Figures 3, 4d

the course exhibited a nice balance between the two, not favoring one over the other. The observed difference of about 0.5 could be attributed to the fact that while all the lectures were comprised of both theoretical and technical parts, none of the assignments had theoretical exercises, which could be perceived as giving a greater emphasis on the technical side, by the students. Among the remaining items, question 4 received the lowest score (which was still significantly better than average), so future implementations

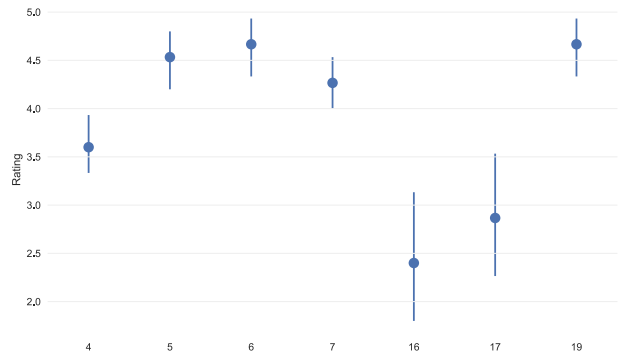

FIGURE 2
Student confidence and course content quality assessment (Questions 4–9, 16–17, 19 from Table 1). Response scale: 1 = Strongly Disagree/Not at all, 3 = Neutral/Somewhat, 5 = Strongly Agree/Very confident.

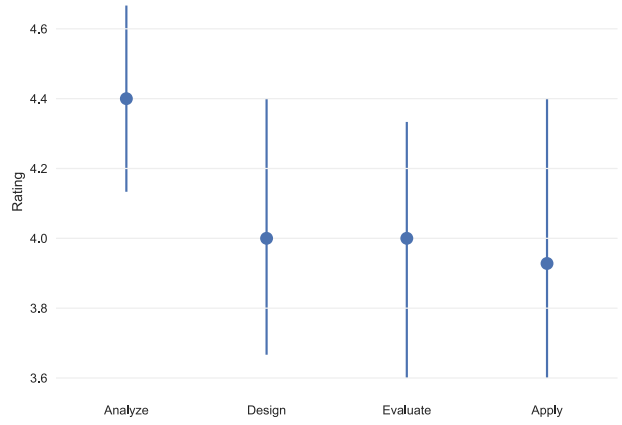

FIGURE 3
Learning objectives achievement assessment (questions 20–23 from Table 1). Response scale: 1 = Unsuccessful, 3 = Somewhat successful, 5 = Successful. Shows student self-assessment of course effectiveness in meeting the four core learning objectives: Analyse, Design, Evaluate, and Apply.

of the course could include more practical aspects that would support students' ability to apply taught XAI techniques in their own models. Apart from that, all other questions received a score significantly higher than 4 (as indicated by the limits of the confidence intervals). This suggested high perceived course effectiveness in building an overall understanding of XAI, while it also indicated that students were satisfied by the diversity of the XAI techniques covered in the course. It is worth noting that question 19, which concerned the overall course satisfaction, received an average score of more than 4.5, with the upper bound of the corresponding confidence interval being very close to 5.

Moving on to Figure 3, all of the objectives received a high score (significantly higher than 3), with *Analyse* having the highest point estimate, 4.4, providing evidence of the students' confidence in deciding the appropriate XAI techniques for the problem at hand. Both *Design* and *Evaluate* achieved a score of 4, suggesting that students felt comfortable designing pipelines for explaining a model and interpreting the final results. Finally, *Apply* had a slightly lower average score, as well as a wider confidence interval, implying there was greater variation in the corresponding scores given by the students. In fact, to get an even more detailed picture of the underlying distributions of scores, Figure 4 shows a collection of barplots representing the relative frequency of each. For all objectives, 4 was the most common score given by students, which indicated their agreement with the statement that the corresponding objective was met by the course. For the *Evaluate* and *Apply*, there was a single student who gave a rating of 2, but otherwise, ratings were mostly on the high end of the spectrum. Having said that, this could be seen as evidence that future implementations of the course would benefit by including additional intermediate assignments, putting more emphasis on the practical aspects of XAI.

To further assess the effectiveness of the course, an analysis of the students' performance on the final project was performed. Each submission was evaluated based on the students' ability to carry out the pipeline shown in the class (data preprocessing, model training, model explanation), as well as the quality of their arguments. Since the project was open-ended, the correction guidelines were that the code should run correctly, the pipeline should be executed reasonably well, and the arguments should be substantial, following the insights gathered from the explanations. Based on that, all students were able to pass the course, demonstrating a sufficient level of competence in performing the aforementioned tasks.

In more detail, about 26% of the students considered both a transparent and a black-box model to address the selected application, although there were no related instructions. This was an indication that (at least a portion of the) students took away the message that black-box models should be used only when achieving significantly better performance than transparent ones. However, since the course was focused on XAI, it was reasonable that most of the students opted for considering just black-box models. Furthermore, about 89% of the students used at least 3 XAI techniques, although 2 was enough to meet the project's requirements. This was an indication that students felt confident to inspect a model from multiple angles, using techniques that bring different insights. Among them, about 87% carried out a comparison between importance scores coming from SHAP and those coming from LIME (Ribeiro et al., 2016), which is another popular XAI technique. This was evidence that students took proactive measures to make sure that a feature's importance was robust. This was in tune with the course material, where it was emphasized that due to the underlying approximations, many XAI techniques suffer from stability issues. Finally, 15% of the students received a borderline pass, due to the fact that although they performed the pipeline adequately, when forming their arguments in favor or against retaining a model, they underutilized the obtained insights. Despite that, their arguments were substantial, so they were sufficient to pass the course, however, they could have been strengthened by taking into account all the available information.

With respect to the qualitative part of the analysis, the open-ended questions revealed specific aspects of the course that

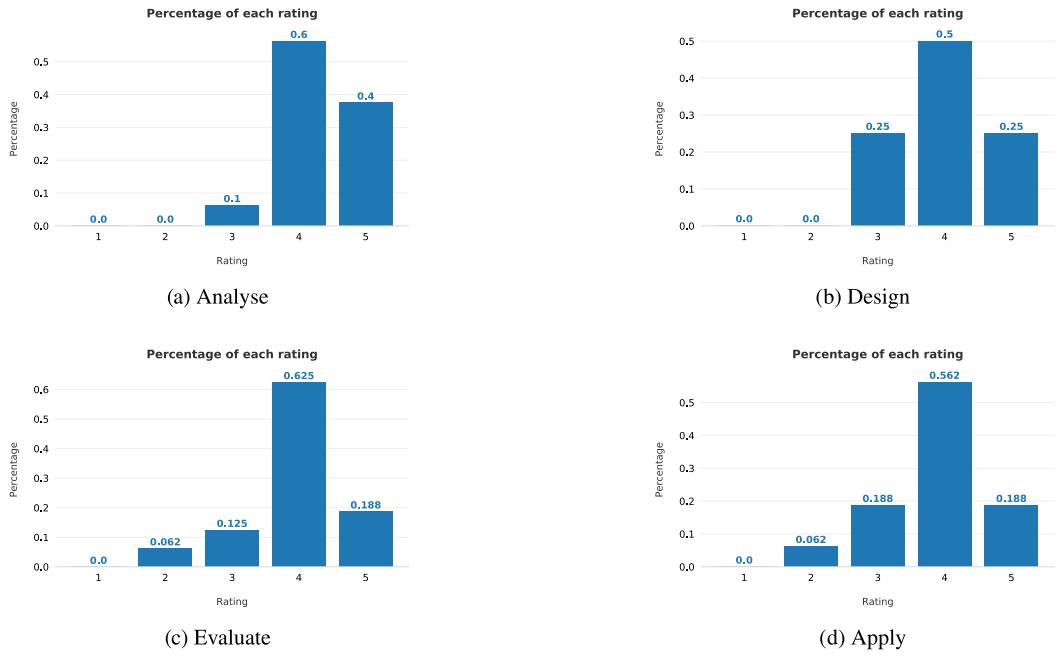

FIGURE 4
Detailed distribution of learning objectives achievement ratings (questions 20–23 from Table 1). Bar charts show the frequency distribution of student responses for each learning objective: **(a)** Analyze: describing ML application context and XAI technique selection, **(b)** Design: defining implementation pipelines, **(c)** Evaluate: critically reflecting on XAI results, **(d)** Apply: competently using XAI techniques and understanding new methods.

were recognized as respondents' favorites. Eleven respondents answered the question "what was your favorite aspect of the course?" Tutorials were mentioned by five respondents. Students appreciated being able to try out the theoretical course aspects in a practical way using the provided workbooks. For example, S-5 said: *"trying out the techniques in the workbooks"*. S-9 said: *"practical application in lab books"*. S-7 said: *"the workbooks and assignment questions. They had the right mix of theory and practical aspects"*. Five respondents mentioned recorded lectures and tutorials. Students appreciated being able to discuss the course material in the online tutorials and lectures. S-1 put it: *"the discussion and Knowledge exchange during Lectures and Tutorial classes"*. S-2 said: *"the meaningful discussions and open-ended questions"*. S-8 said: *"the ambience of the teams' sessions is done with a "brainstorming" approach which gives us the opportunity to discuss ideas, bring questions from the real world and hear different opinions"*. Respondents also mentioned open-ended questions, sufficient examples, and assignment questions. Overall responses were very positive, for example, S-2 said: *"I was very impressed with the structure and delivery of the material...It [the course] made me not only appreciate the XAI fundamentals but the whole approach toward applying ML algorithms...I consider myself very lucky for selecting this course and I believe it has helped me tremendously in my understanding of ML projects."* S-6 said: *"I have understood why the area of XAI techniques has gotten attention and is important to make AI / ML available for general use in the Data Analytics Project."* S-8 reflected: *"The topic of XAI it's very interesting! Thank you for including it in the program and giving us exposure to these approaches."*

5 Discussion and implications

Overall, the survey results supported the claim that the course material and its delivery can be highly effective in teaching XAI techniques. Analysis of individual ratings showed that this course was especially useful in promoting understanding of a diverse array of XAI techniques, and their conceptual distinctions, advantages, and disadvantages. The respondents' answers to open-ended questions suggested that interactive and practical aspects of the course were important in the successful process of translating XAI theory into practical skills. The open questions and codebooks were also important parts of the course, especially in combination with the ensuing discussions. The survey results also suggested that the course can be effective in teaching XAI techniques to individuals having no or minimal experience and knowledge about XAI.

Our experience suggests several key principles for successful XAI education. First, adopt a narrative-driven approach that organizes content around explanation types and stakeholder questions rather than chronological technique development. This helps students understand when and why to apply specific techniques rather than simply learning isolated tools. Second, emphasize multi-disciplinary connections throughout the course. Link SHAP to game theory, counterfactuals to philosophy and causal inference, and all techniques to ethical AI considerations. This broader context helps students appreciate XAI as an interdisciplinary field rather than a purely technical domain. Third, maintain an application-focused assessment philosophy that prioritizes interpretation and practical application over

mathematical derivations. While theoretical understanding is important, the primary goal should be developing competent practitioners who can appropriately deploy XAI techniques in real-world contexts.

Students' performance and forum questions suggested that, besides the technical aspects, such as consideration of Python libraries updates, the course could be improved by providing more support for the output analysis and evaluation aspect of the XAI. Most of the students found this part most challenging. This was also reflected in a slightly lower overall score of the evaluation learning objective, i.e., the ability to critically reflect on the results from the XAI techniques and suggest how it helps the given context. This could be because the selected datasets were not relevant to students' professional or research interests, however, it could also mean that more practical exercises focused on the analysis part should be included in the course. Potentially, more practice analyzing XAI outputs could lead to a better understanding.

To strengthen XAI output interpretation skills, we recommend implementing structured interpretation exercises within each tutorial. These include: (1) Guided Analysis Templates that provide step-by-step frameworks for interpreting SHAP plots, counterfactual results, and anchor rules, with explicit questions like "What does a negative SHAP value indicate about this feature's contribution?" and "How would you explain this counterfactual change to a loan officer?" (2) Comparative Interpretation Assignments, where students analyse the same dataset using two different XAI techniques and explain discrepancies in their insights; and (3) Error Analysis Exercises, where students are given intentionally problematic XAI outputs (e.g., SHAP explanations from biased models) and must identify potential issues and limitations. Additionally, we suggest incorporating interpretation checkpoints throughout tutorials where students must pause and explain what specific visualizations or outputs mean before proceeding to the next step.

Finally, build flexibility into the framework to accommodate different instructor preferences and student backgrounds. The modular structure allows technique substitution (e.g., replacing Anchors with LIME) while maintaining the overall pedagogical approach. Begin each course by establishing the transparent versus opaque model distinction to motivate XAI necessity, use consistent datasets across techniques to enable meaningful comparisons, and require final projects that demonstrate integration of multiple techniques for comprehensive model evaluation. This framework emphasizes developing principled understanding that will transfer to new XAI techniques as the field evolves, while ensuring students gain practical experience with current state-of-the-art methods.

5.1 Limitations

The scores of the questionnaire were self-reported and reflected the subjective evaluation of respondents' own understanding of the course material. The high evaluations of the course effectiveness were reflected in the objective assignment performances. Although this course has been delivered to data science experts working within the banking sector, in this chapter, only the students' responses were analyzed. This limitation prevented evaluation of the generalizability of the course effectiveness across different settings and expertise levels. In the future, further surveys will be conducted to assess the effectiveness of the course for the more experienced data science and ML experts strictly working in a professional setting.

6 Conclusions

This work presents an approach for structuring a course on explainability in machine learning. The aim of the course was to provide a formal introduction to the field of explainability. Although advances in XAI come at a rapid pace, the fundamental ideas and objectives are likely to remain the same. The course was primarily designed for industry professionals, data scientists, and students with a programming and data science background. One of the main drivers governing the development of the material was to address the reported misuse of XAI techniques in practical applications.

To this end, a combination of diverse XAI techniques was included in the material, focusing on conceptual details and distinctions between different explanation types. Finally, most of the lectures were accompanied by a workbook demonstrating the function and utility of the corresponding XAI technique, in order to allow students to gain some hands-on experience. Students' feedback and performance provided strong evidence that the course was effective in meeting its learning objectives.

Data availability statement

The datasets presented in this article are not readily available because the generated dataset is for internal university purposes, such as identifying student complaints and improving course quality. Requests to access the datasets should be directed to Ioannis Papantonis, gppntonis@gmail.com.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent from the participants was not required to participate in this study in accordance with the national legislation and the institutional requirements.

Author contributions

AB: Software, Writing – review & editing, Methodology, Writing – original draft, Conceptualization, Formal analysis. IP: Formal analysis, Methodology, Writing – review & editing, Software, Writing – original draft. AS: Methodology, Writing – review & editing, Formal analysis, Writing – original draft. VB: Writing – review & editing, Supervision, Writing – original draft.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This research was partly supported by a Royal Society University Research Fellowship, UK and partly supported by a grant from the UKRI Strategic Priorities Fund, UK to the UKRI Research Node on Trustworthy Autonomous Systems Governance and Regulation (EP/V026607/1, 2020-2024).

Acknowledgments

We would like to thank all the students in the course for their effort, dedication, as well as their valuable feedback.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2025.1595209/full#supplementary-material

References

Arrieta, A. B., DÃaz-RodrÃguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., et al. (2019). Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* 58, 82–115. doi: 10.1016/j.inffus.2019.12.012

Arya, V., Bellamy, R., Chen, P.-Y., Dhurandhar, A., Hind, M., Hoffman, S., et al. (2019). One explanation does not fit all: a toolkit and taxonomy of AI explainability techniques. *arXiv preprint* arXiv:1909.03012. doi: 10.48550/arXiv.1909.03012

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE* 10:e0130140. doi: 10.1371/journal.pone.0130140

Belle, V., and Papantonis, I. (2020). Principles and practice of explainable machine learning. *Front. Big Data* 4:688969. doi: 10.3389/fdata.2021.688969

Bennetot, A., Donadello, I., Qadi, A. E., Dragoni, M., Frossard, T., Wagner, B., et al. (2021). A practical tutorial on explainable ai techniques. *ACM Comput. Surv.* 57, 1–44. doi: 10.1145/3670685

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics* 19, 15–18. doi: 10.1080/00401706.1977.10489493

Deng, H., Guan, X., and Khotilovich, V. (2014). "inTrees: Interpret Tree Ensembles [dataset]," in *CRAN: Contributed Packages* (The R Foundation). doi: 10.32614/CRAN.package.inTrees

Efron, B., and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statist. Sci.* 1:815. doi: 10.1214/ss/1177013815

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Ann. Stat.* 29:1013203451. doi: 10.1214/aos/1013203451

Goldstein, A., Kapelner, A., Bleich, J., and Pitkin, E. (2013). Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation. *J. Comput. Graph. Stat.* 24, 44–65. doi: 10.1080/10618600.2014.907095

Griffin, P., Coates, H., Mcinnis, C., and James, R. (2003). The development of an extended course experience questionnaire. *Quality High. Educ.* 9, 259–266. doi: 10.1080/135383203200015111

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., and Wortman Vaughan, J. (2020). "Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. doi: 10.1145/3313831.3376219

Kingma, D. P., and Ba, J. (2015). Adam: a method for stochastic optimization. *arXiv:1412.6980*.

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., et al. (2016). *Jupyter notebooks-a publishing format for reproducible computational workflows*. Technical report, IOS Press, 87–90. doi: 10.3233/978-1-61499-649-1-87

Koh, P. W., and Liang, P. (2017). Understanding black-box predictions via influence functions. *arXiv:1703.04730*.

Lakkaraju, H., and Lage, I. (2019). "Interpretability and explainability in machine learning," in *COMPSCI 282BR*.

Lewis, D. (1973). Causation. *J. Philos.* 70:556. doi: 10.2307/2025310

Lundberg, S. M., and Lee, S.-I. (2017). "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17* (Red Hook, NY, USA: Curran Associates Inc.), 4768–4777.

McInnis, C. (1997). Defining and assessing the student experience in the quality management process. *Tert. Educ. Manag.* 3, 63–71. doi: 10.1080/13583883.1997.9966908

Miller, T. (2019). Explanation in artificial intelligence: insights from the social sciences. *Artif. Intell.* 267, 1–38. doi: 10.1016/j.artint.2018.07.007

Ramsden, P. (1991). A performance indicator of teaching quality in higher education: the course experience questionnaire. *Stud. High. Educ.* 16, 129–150. doi: 10.1080/03075079112331382944

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why should i trust you?," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. doi: 10.1145/2939672.2939778

Ribeiro, M. T., Singh, S., and Guestrin, C. (2018). Anchors: high-precision model-agnostic explanations. *Proc. AAAI Conf. Artif. Intell.* 32:11491. doi: 10.1609/aaai.v32i1.11491

Rothman, D. (2020). *Hands-on Explainable AI (XAI) with Python: Interpret, Visualize, Explain, and Integrate Reliable AI for Fair, Secure, and Trustworthy AI Apps*. Birmingham: Packt Publishing, Limited.

Shapley, L. S. (1952). *A Value for n-Person Games*. Santa Monica: The Rand Corporation.

Wachter, S., Mittelstadt, B., and Russell, C. (2018). Counterfactual explanations without opening the black box: automated decisions and the GDPR. *Harvard J. Law Technol.* 31, 841–887. doi: 10.2139/ssrn.3063289