TXAI Practical Exercises – Collected Answers

Miguel Leal Fernandez (miguel.leal@rai.usc.es), Gian Paolo Bulleddu (gianpaolo.bulleddu@rai.usc.es)

Contents

1 Introduction

This document contains the answers to the exercises from Laboratory 1 and Laboratory 2 of the Explainable and Trustworthy Artificial Intelligence (TXAI) course, 2026. For each laboratory notebook, a short summary of the activities performed is first provided, then the list of exercises and the realted answers follows.

2 Notebook I1 – Introduction to SHAP

2.1 Summary of the notebook

This notebook introduces SHAP (SHapley Additive exPlanations) as a method to interpret machine learning models. The workflow includes loading the diabetes dataset, training several models such as Decision Trees and Random Forests, visualizing the trained models, and generating both global and local SHAP explanations. Different model configurations are compared to analyze how model structure affects feature importance and prediction explanations.

2.2 Exercises and Answers

Exercise I1.1

Try to build some alternate versions of the Decision Trees we obtained, by changing the parameters on the constructor call. You don't need to be exhaustive: trying two or three is enough. You may keep limiting their depth to keep their interpretation manageable . Plot these new decision trees and compare them to the two trees (**dtc** and **dtc5**) given as illustrative examples in the previous cells. Are they significantly different? Extract some conclusions as to why/why not, and summarize a few key insights onto the diabetes prediction problem you can distill from the trends shown by your set of trees.

In this exercise we have built two decision tree classifiers by variyng some hyperparameters. In both models we have kept the maximum depth to 5 then in the first model we have chamged the Gini impurity criterion to Entropy while in the second model we have set the minimum number of leaf to three. The purpose of this task is to evaluate how the decision logic differs from the predefined decision trees. After the training of the newly created models we have analyzed the results and down below are our conclusion.

First of all we have not detected big structure difference between the new (**dtcv1**,**dtcv2**) and the original trees (**dtc**, **dtc5**). The 'plasma glucose concentration' feature most frequently appears as the root node or very close to it in all models, this suggest that this is the most important feature. The 'Body mass index' too is often selected in the upper levels of the trees frequently acting as a secondary decision variable . 'Age'and 'diabetes pedigree' are most frequently selected at an intermediate tree level. The rest of the dataset features like 'insulin','diastolic blood pressure','triceps skin fold thickness' have a tendency to appear only in the deeper levels of the trees or not at all so we can conclude that those are the less important features.

The Entropy splitting criterion has not lead the model to big changes in tree structures and split thresholds, moreover it has not changed the importance or the ordering of the dataset features. The entropy-based tree selects similar variables at the top of the tree and produces comparable decision paths,regardless of the impurity criterion used the the most important features can be identified by the models.

Limiting the minimum number of sample per leaf produces simpler trees , however even reducing highly specific leaves and eliminating splits based on few data observations the structure of the tree looks almost unchanged, especially at tree top levels. The main difference in this model tree does not have any node branched by the feature ‘times being pregnant‘.

In conclusion all trained trees don’t show big differences in structures and logic, the prediction of diabetes diagnosis looks mainly driven by features ‘plasma glucose concentration‘ and ‘Body mass index‘,‘Age‘and ‘diabetes pedigree‘ are important as well but with a lower strenght in routing the model to a good prediction.

Exercise I1.2

Do the SHAP values for the Decision Trees match your inspection of their graphical representation?

Yes, the SHAP values confirm what we saw in the decision tree. Plasma glucose concentration has the highest impact on the predictions, just like it appears at the top of the tree splits. Other features, such as body mass index, age, and diabetes pedigree function, also show significant influence, which matches their presence in the higher levels of the tree. The SHAP summary plot shows that high values (red points) of plasma glucose generally push predictions toward the positive class (dots on the right), consistent with the tree’s decision rules.

Do you find significant differences between the Decision Trees and the Random Forest?

Yes, there are differences between the two types of models. The Random Forest exhibits less dispersion in SHAP values, with points more densely grouped and fewer outliers. This shows the ensemble nature of the Random Forest, where predictions are averaged over many trees, reducing variability and sensitivity to individual data points. In contrast, single Decision Trees tend to produce more scattered SHAP values and sharper transitions due to their reliance on hard decision thresholds.

Exercise I1.3

Try your hand at generating new plots (both global and local) following the code given during the section.

Do you think they offer better information than the summaries we were using? Which plotting options do you think would be the best at explaining the problem at hand to a layperson?

Would your answer change if the explanations were meant for an expert?

Yes, these plots do offer better and more complete information than the summary plot alone, however it has to be seen as a complementary rather than substitutive way. The SHAP summary plot helps in understandin the global behavior of the model, it highlights which the most importance features are and how their values influence the predictions. However, it is quite abstract and does not explain individual model decisions.

The bar and waterfall plots provide clearer local explanations. The bar plot makes it easy to see which features contribute the most to a single prediction, while starting from a baseline ,the waterfall plot explicitally shows how each feature influences the model’s output toward or away from the predicted class. This makes the decision process more interpretable at the instance level and helps understand why a specific patient is classified as diabetic or not. Partial dependence plots add another perspective by showing how a feature affects predictions on average across the dataset, which can be useful to understand general trends.

For a a person without professional or specialized knowledge in this particular subject, the bar and especially the waterfall plots are the most effective. They are intuitive, visually clear, and allow to easily explain which factors increase the risk and which decrease it for a given individual. In contrast, the summary plot and partial dependence plots can be harder to interpret without technical background.

If the explanations were meant for an expert, the answer would change. An expert audience would benefit more from the summary plot and partial dependence plots, as these provide global insights into model behavior, feature interactions, and overall consistency. In that case, local plots would still be useful, but mainly as a complement for analyzing specific cases or debugging the model rather than as the primary explanation tool.

Exercise I1.4

With the Songs dataset imported, follow the process laid out through the notebook to explain the **global** behaviour of the model. Then, find a song you like (and another song that you don't like) on the dataset (using the code provided), and **locally explain** the predictions made by a decision tree with a good interpretability-accuracy trade-off and Random Forest. Discuss all your findings.

Global explanation: In analyzing the decision tree we have observed that the feature ‘energy‘ is the root of the tree and drives the first split so that we can coclude that it is the most important feature for the model. This fact can suggest that the model distinguishes songs depending on how much energetic they are. After ‘energy‘, even ‘tempo‘ and ‘loudness‘ features play an important role in the model decision logic, while features such as ‘danceability‘ and ‘duration‘ are used to refine the decisions at deeper levels of the tree, these last features contribute to the final classification but are less importsnt than ‘energy‘, ‘tempo‘, and ‘loudness‘. The SHAP summary plot provides a complementary point of view. According to SHAP plot, ‘instrumentalness‘ and ‘loudness‘ are the most important features . In particular, for ‘loudness‘, it is clear that higher values tend to force the prediction toward the “Dislike” class, very loud songs are generally less preferred by the model. Moreover, the SHAP plot shows some outliers related to ‘energy‘, where very low ‘energy‘ values strongly contribute to a “Dislike” prediction. This shows that, even if ‘energy‘ is an important feature in the tree's decisions, very high or very low energy values can strongly affect the model's predictions. When comparing the SHAP summary plots of the Decision Tree and the Random Forest, we can notice small differences. Since the Random Forest combines the predictions of many trees, it produces a smoother distribution of SHAP values, which reduces variability and the influence of individual splits. In contrast, the Decision Tree assigns zero or near-zero importance to some features, since a single tree may never use them in its splits. This explains why certain features appear unimportant in the Decision Tree SHAP plot but still have importance in the Random Forest. In the end,both models are able to identify the same factors that influence musical preferences,the Random Forest provides a more reliable and detailed explanation ,while the Decison Tree provides simpler and easier to understand decision rules.

Local explanation: The selected songs for local analysis are Beat It by Michael Jackson (ID 1828 in the original dataset, corresponding to index 228 in the test set) and Hips Don't Lie by Shakira (ID 1927, corresponding to index 327 in the test set). Both songs were analyzed using the Decision Tree and the Random Forest in order to compare their local explanations. For Beat It, the local SHAP explanations show that loudness and instrumentalness are the most influential features, and both contribute negatively to the probability of the song being liked. This suggests that, for this specific instance, high loudness levels and low instrumental content push the prediction toward the “Dislike” class. On the other hand, danceability has a positive contribution, partially compensating for the negative effect of loudness and instrumentalness and supporting the “Like” prediction. This combination of features reflects a trade-off between rhythmic appeal and production characteristics in the model's decision. In the case of Hips Don't Lie, the same three features—loudness, danceability, and instrumentalness—also appear as the most important. However, their effects differ depending on the model. In the Decision Tree, loudness contributes positively to the prediction, indicating that higher loudness levels increase the likelihood of the song being liked for this instance. When analyzing the Random Forest, the importance of loudness and danceability decreases, while instrumentalness becomes the dominant factor by a large margin. This change reflects the more stable and averaged behavior of the Random Forest, which smooths the influence of individual features and relies more heavily on consistent patterns across many trees. Overall, the local explanations highlight how the same features can have different impacts depending on both the specific song and the model used, emphasizing the value of local interpretability tools such as SHAP.

3 Notebook I2 – Building Explanations in Natural Language

3.1 Summary of the notebook

This notebook introduces fuzzy inference systems (FIS) as interpretable predictive models capable of producing linguistic explanations. Using the IRIS dataset, different fuzzy systems are generated with alternative partitioning and simplification strategies. The notebook compares rule bases, prediction behavior, and interpretability–accuracy tradeoffs, including analysis through Pareto fronts.

3.2 Exercises and Answers

Exercise I2.1

Observe the system we ended up with:

It seems that we "lost" two fuzzy variables, considering that IRIS has 4 features. This is obviously an effect of simplification, but why did simplification take those two out?

The simplification process removed variables that contributed little discriminative information relative to Petal Length and Petal Width. Because these two features already separate the classes effectively, retaining the remaining variables would only increase model complexity without significantly improving predictive performance. Therefore, the simplification step is justified as it reduces redundancy while preserving classification quality.

Consider the data distributions you see on the histograms for Petal Length and Petal Width. Then, consider the fuzzy partitions shown for those two variables. What do you think the system did to simplify the partitions for Length? Do you think that whatever it is, it is justified in light of the underlying data distribution?

The system simplified the fuzzy partitions for Petal Length by reducing the number of distinct fuzzy sets and merging regions where different classes exhibit similar values. This happens because Petal Length shows overlapping distributions for two of the target classes, which makes it difficult to justify highly granular partitions. As observed in the pair plots involving Petal Length, classes 2 and 3 form a single, dense cluster, while class 1 is clearly separated. As a result, the system creates a broader fuzzy region that captures the overlapping values of classes 2 and 3, corresponding to an "average" or "medium" linguistic term, while assigning a separate fuzzy set (such as "high") to class 1. This simplification is justified by the underlying data distribution, as introducing more detailed partitions for Petal Length would not meaningfully improve class separation and would instead add unnecessary complexity. By aligning the fuzzy partitions with the natural clustering of the data, the system achieves a good balance between accuracy and interpretability.

Exercise I2.2

Compare the features of these two new systems against the one we took as baseline.

What differences appear in each case? Why?

When comparing the two alternative systems with the baseline system (RP_FDTP_S), several differences can be observed, mainly related to feature selection and fuzzy partitioning. The system RP_FDTP uses all the original features of the IRIS dataset, each one divided into three regular fuzzy partitions (low, average, and high). This is because no simplification is applied, neither at the feature level nor at the partition level. As a consequence, the rule base is larger and more complex, even though some of the rules are effectively redundant. In fact, we can observe that rules 3, 4, and 5 in RP_FDTP correspond closely to rules 2, 3, and 4 in the baseline system, indicating that the additional variables do not provide new discriminative information but rather replicate patterns already captured by the most informative features. The system SP_FDTP_S, on the other hand, applies simplification while using induced (data-driven) partitions. As a result, it retains the same relevant attributes as the baseline system, namely Petal Width and Petal Length, confirming their importance for classification. However, the fuzzy partitions differ: while the baseline system uses three partitions for both variables, SP_FDTP_S uses two partitions for one feature and three for the other. Additionally, the three-partition variable is no longer symmetrically divided, reflecting the actual data distribution more closely. This asymmetry suggests that the system adapts the fuzzy sets to the natural clustering present in the data rather than enforcing uniform linguistic terms.Overall, these differences arise from how each system balances interpretability, complexity, and fidelity to the underlying data distribution.

If you had to choose only one alternative, without further information, which would you choose? Why?

Without further information, the system SP_FDTP_S would be the preferred choice. This system combines pruning and simplification with induced fuzzy partitions, allowing it to adapt the fuzzy sets to the actual data distribution rather than relying on uniform, predefined partitions. As a result, it captures the underlying structure of the data more accurately while still maintaining a compact and interpretable rule base. Compared to RP_FDTP, SP_FDTP_S avoids unnecessary complexity and redundancy by discarding irrelevant features and overly detailed partitions. When compared to the baseline RP_FDTP_S, the induced partitions provide additional flexibility, especially in cases where class distributions overlap or are unevenly distributed. Although the resulting fuzzy sets may be less symmetric, they better reflect real data behavior and do not significantly reduce interpretability. Overall, SP_FDTP_S offers a better balance between expressiveness and simplicity by aligning model structure with the data itself, making it the most robust choice given the available information.

Exercise I2.3

Taking the code given as an example, replicate the inference using the alternative systems defined above (no simplification / induced partitions). Observe and discuss the results obtained.

Observe and discuss the results obtained.

For instance 145, all three fuzzy inference systems correctly classify the sample as class 3 (Virginica). This confirms that, despite their structural differences, all systems are able to capture the relevant decision logic for this instance. However, the way in which each system reaches the final decision differs, which highlights the explainable-by-design nature of fuzzy models. In the baseline system (RP_FDTP_S), multiple rules are activated simultaneously with different firing strengths. One rule weakly supports class 2, while two other rules support class 3, with the strongest contribution coming from the rule based solely on Petal Width being high. This illustrates how competing rules act as competing explanations, allowing us to see not only the final decision but also alternative classifications that were considered and ultimately outweighed. The RP_FDTP system, which removes simplification but keeps regular partitions, produces an almost identical explanation to the baseline. The same attributes are involved, and the same rules fire with the same strengths, although they are indexed differently due to the presence of additional unused variables. This indicates that removing simplification does not significantly affect either the prediction or its explanation for this instance, as the extra features do not meaningfully contribute to the decision. In contrast, the SP_FDTP_S system, which uses induced partitions and simplification, produces a noticeably different explanation. Only two rules fire, both of which directly support class 3. The decision is driven by broader and asymmetric fuzzy sets that better align with the data distribution, leading to higher firing strengths and a more compact rule base. Although the final prediction is the same, the reasoning process is simpler and more decisive, with no competing rules supporting alternative classes.

Which of the two modifications created a higher deviation from the baseline?

Systems using induced partitions tend to deviate more from the baseline because they adapt fuzzy sets to the empirical distribution of the data. This flexibility can improve representation fidelity but may slightly change prediction outcomes compared to systems based on fixed partitions.

Exercise I2.4

Check the given Pareto Front:

Why do you think the FIS may be placing above the DT in the tradeoff?

The Fuzzy Inference Systems place above the Decision Trees in the interpretability-accuracy tradeoff because they achieve comparable or even higher accuracy with a significantly lower structural complexity. While Decision Trees rely on many hard splits and leaves to model the decision boundaries, FISs use a small number of fuzzy rules that can cover broader regions of the feature space thanks to fuzzy memberships. This allows them to represent complex decision logic with fewer rules. In addition, fuzzy rules are closer to natural language (like saying “Petal Width is high”), which improves semantic interpretability compared to the numerical thresholds used in Decision Trees. As a result, FISs can maintain high accuracy while remaining compact and easier to understand, which explains why they appear in a better position on the Pareto Front.

Induced Partitions seems to be slightly better than Regular Partitions, at least in light of the results shown. Can you imagine a situation in which, with this same Pareto Front, the Regular Partitions alternative would be preferred? Why?

Yes, there are situations in which the Regular Partitions alternative would still be preferred, even if induced partitions show slightly better accuracy on the Pareto Front. Regular partitions produce symmetric and uniformly distributed fuzzy sets, which are often easier to interpret and communicate, especially to non-expert users. When interpretability, transparency, and ease of explanation are more important than marginal gains in accuracy, regular partitions may be favored. Moreover, regular partitions are less dependent on the specific data distribution and may generalize better when the dataset is small, noisy, or subject to change. In scenarios where robustness and stability are prioritized over fine-grained adaptation to the training data, regular partitions can reduce the risk of overfitting introduced by induced, data-driven partitions. Therefore, despite slightly lower performance, regular partitions may be preferred in applications where simplicity, consistency, and trustworthiness are critical.

Exercise I2.5

After define, import and predict/explain the PIMA dataset

Would you prefer this alternative to the SHAP/Visual approaches we used on the previous practical? Why?

The fuzzy inference system produces explanations that are expressed in a language similar to human language, making them easy to understand and interpret.

However, this approach show its limitations when the behaviour of complex machine learning models needs to be explained and interpreted. SHAP explainations are less intuitive and more difficult to undeerstand but they are more flexible and generally provides a more accurate estimations of the feature contributions.

Finally, we would prefer fuzzy systems when interpretability is our main objective, while SHAP approaches would be preferred for explaining complex machine learning models.