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 Artificial Intelligence (XTAI) course, Spring 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.

1.1 Note on the formulation of the questions

Some of the original exercise questions have been slightly rephrased in this document to adapt them to the present written format and to ensure clarity outside the original notebook context. The meaning and intent of the questions, however, remain unchanged with respect to the original laboratory notebooks.

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.

The same modeling and explanation pipeline was applied to the SONGS dataset. After training a classifier, SHAP explanations were computed to identify the features most strongly associated with each prediction. The analysis confirmed that the explanation framework remains applicable across datasets and helps identify the most influential musical attributes driving the classification outcomes.

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?

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 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.

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?

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

The RP_FDTP system uses all original features with regular fuzzy partitions, leading to a larger and more complex rule base, often containing redundant rules. The SP_FDTP_S system applies simplification and induced partitions, retaining only the most relevant attributes while adapting partitions to the data distribution. Without further information, SP_FDTP_S would be preferred because it balances interpretability and predictive accuracy, maintaining compactness while reflecting the underlying data structure more faithfully.

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.

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?

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?

Fuzzy systems can achieve a favorable interpretability–accuracy balance because they use compact rule bases expressed in linguistic terms while still capturing nonlinear decision boundaries. Regular partitions might be preferred in settings where interpretability consistency or linguistic symmetry is more important than small predictive improvements.

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?

A new FIS was generated using the GUAJE tool and applied to the PIMA dataset. The resulting system produced interpretable rules linking clinical variables with diabetes outcomes, demonstrating how fuzzy modeling can generate both predictions and natural language explanations.