# XAI Lab Exercises

Questions

## Lab 1 - Intrinsic Explainable Models

**Q 1)** Familiarise yourself with the libraries.

- **a)** Read the description of the libraries that are used and what is the purpose of each in the import cell.

**Q 2)** Load the Data

- **a)** Add the name of the file and it's format.
- **b)** Find the method in "pandas" libraries to read this file forma. Reference to `pandas` io submodule
- **c)** Explain the following default arguments for the read method and how they link to the data being imported:
  - i. Explain the use of 'sep' or 'delimiter' argument.
  - ii. Explain the use of 'header' argument
  - iii. Explain the use of 'decimal' argument and its importance in financial data sets.

  Note: to be able to explain AI methods, it is also important to have good data analyst skills, since if the data we use can not be explained or is imported incorrectly, we are then trying to explain meaningless results.

**Q 3)** Perform Basic Exploratory Data Analysis

For our exercise we want to explain the classification from the first 13 attributes to predict the target class, if there is a heart disease or not.

- **a)** From your theory explain which model(s) can be used for a classification task.
- **b)** If for some reason you were limited to using only a few attributes, for instance choose 5 out of 13 attributes from the data set to train a model, how would you choose those attributes? (note: we can not use any advanced XAI methods yet since they require a trained model)

  HINT: this method's result can be shown using a heatmap.
- **c)** How does the answer the previous question assist the explanation of the model?
- **d)** Perform the method in your answer to question 3) b). Use `matplotlib.pyplot.figure()` for creating a figure and appropriate method in pandas library for the stated method's calculation then use `seaborn.heatmap(*arguments)` with appropriate arguments to display the results of the calculation.

  `MatPlotLib.PyPlot.Figure`

  `Pandas Library Method`

  `Seaborn Heatmap`
- **e)** What are the top 5 most "explainable" attributes for the target class?

**Q 4)** Using Intrinsic Explainable Models

- **a)** Train a Logistic, Tree, and kNN model on the training data.
- **b)** List the parameters that were fit in the logistic regression model and write a brief explanation on what these parameters represent in a logistic model.

  Hint: Read the article
- **c)** Produce a decision tree from the trained model something that a person can use to explain a decision of the model.

**d)** Explain the importance of different parameters values in Decision Trees and $k$NN models.

**Q 5)** Intrinsic Model Experiments

**a)** Use different methods of splitting for Decision Tree. Plot the different accuracies and comment on how each split is different.

**b)** Plot the accuracy scores for different values of $k$ in the $k$NN model and comment on how the explanation changes with different values of $k$.

## Lab 2 - Model Agnostic Methods

**Q 1)** Permutation Importance

**a)** Perform Permutation Importance for logistic model for all the attributes and plot the results. Give a brief interpretation of your results.

Reference: `scikit.inspection`

**b)** Explain which model is most likely to be show a bigger difference in prediction when using permutation importance.

**c)** Perform Permutation Importance for two features.

i. Choose the most explainable features and motivate why you have chosen these two features.

ii. Run the permutation importance for each model. Comment on the differences.

Reference to submodule `sklearn.inspection.PartialDependenceDisplay`

**Q 2)** KernelSHAP

Note: You will need to have the `jupyter` notebook to be trusted to use the `javascript` methods.

**a)** Use a SHAP Explainer to derive SHAP Values for the logistic regression model.

**b)** Plot a SHAP summary plot using all the features in the data and explain the results of the plots.

**c)** Plot the partial dependence plot using the SHAP values for the attribute `exang` and explain how to interpret the plot.

**d)** Plot a `force plot` for a random $X_test$ data point, and explain the figure.

**e)** Plot the summary plot for each of the attributes and give an interpretation of the plot.

(You will need to have the `jupyter` notebook to be trusted to use the `javascript` methods.)

Reference: `shap`

**Q 3)** Dependence Plot

**a)** Plot a dependence plot to show the effect of 'chol' across the whole dataset.

**b)** Plot the two-way PDP showing interactions between features 'Resting blood pressure' and 'Chest pain type' and explain their effect.

**Q 4)** Perform the same analysis as above for a model of your choice (other than the Logistic Model) and comment on the sensitivity of the two models by comparing and contrasting the results.

## Lab 3 - Explaining Deep Learning Models

**Q 1)** Occlusion

**a)** Use the provided starter code to generate areas of importance in classification of random images.

*Use the Heatmap to explain qualitatively how well the pre-trained model is able to understand the object of the image.*

We have given a running example for a basketball image, but continue your exploration for all other images in the directory.

**b)** Explain what is important in chosen data set, so that what we as humans intuitively use to identify objects, can be learnt by the machine.

Hint: refer back to lectures with example of on mis-classification.

**Q 2)** Unsupervised Explainers

Note: This does not require a deep learning model.

**a)** Using PCA find the number of dimensions (axis) with the most variance (i.e. the components have at least 0.1 variance).

**b)** Perform PCA and t-SNE on the Fashion-MNIST dataset using the given the starter code.

**c)** Explain what can be interpreted using an unsupervised method like this in explaining an incorrect classification.

**Q 3)** Gradient Based Explainers

**a)** Use the given data set and pretrained model to generate heatmap overlay for which part of the image explains the classification, i.e. perform a CAM for a random image.

*A running example of aircraft carrier and submarine is given*

**b)** Explain how when there are multiple objects in a frame, this could be used identify each object individually.

**Q 4)** DeepSHAP

**a)** Use the shap library and the provided starter code to explain the MNIST data set.

**b)** Explain what can be understood about the similarity of certain digits as seen by the machine learning model through the explanations.

**Q 5)** Attacks and Defence using Adverserial Features

**a)** Read the article NeurIPS 2017 - Adversarial Attacks and Defences Competition, mainly Section 2.2 and 2.3 and explain the general idea what the attack and defence try to achieve. (Explain at least 2 different of each attack and defence)