Course abstract:
AI systems are very good at making predictions in a variety of settings. In many cases, however, this comes at the expense of interpretability of the models used. In this course, we will see how to interpret the decisions made by these systems such that they are accessible to humans. We will look at a number of practical things that can be done to improve an algorithm's interpretability and explainability.

We will provide an overview of different methods for interpreting classifiers. We will look at machine learning models which are interpretable by nature as well as model-agnostic methods for interpreting classifiers. We will cover feature importance, partial dependence plots, local interpretable model-agnostic explanations, and Shapley additive explanations.

This course is both theoretical and practical.

Practical elements:
The practical exercises will be done in Python. We will implement several classification algorithms and analyse how interpretable they are. We will analyse how features contribute to the classifications. We will also use methods that are able to explain black box models to see how they make predictions.

Prerequisites/knowledge:
Basic knowledge of Python would be good.

Will I be using particular software?
Python3 on Google Colab platform (requires a gmail account only).

Will I need to bring a laptop, or download anything in advance?
No.

Will I need to prepare anything in advance?
If bringing a laptop, participants will need Python3, Jupyter and the following libraries: scikit-learn, keras, shap, lime, numpy, pandas, graphviz, pydotplus.