Assignment: Phishing web sites

Learning goals

In this assignment, you:

1. learn to build a decision tree classifier.
2. improve your data manipulation skills in Python.

Assignment

Phishing refers to a family of online frauds where an Internet user is lured into submitting his/her sensitive data for malicious purposes.

Load the phishing data set from either location:

- **Documents/Methods/Data/Phishing** folder in the course's Oma workspace (preformatted into CSV for convenience)
- https://archive.ics.uci.edu/ml/machine-learning-databases/00327/Training%20Dataset.arff (in Arff format; easily modifiable into CSV).

Data source: https://archive.ics.uci.edu/ml/datasets/phishing+websites

Note: As the interpretation of the -1's and 1's in the Result column seems to be missing from the document, it may be helpful to know that a '1' corresponds to a phishing site and a '-1' to a legitimate site.

Your goal is to construct a small yet useful decision tree that predicts whether a website is a phishing site or not.

Your result must contain all of the following:

1. An image of the final decision tree.
2. Written instructions for an internet analyst to make the decision of whether the website is likely to be a phishing site or not. The instructions must match one-to-one with your decision tree.
3. The accuracy estimate (percentage of correct classifications) of your decision tree.
4. A copy of the Python code used.

Hint

An important thing to note: although decision tree classifiers are designed for categorical data, the **sklearn** implementation requires the explanatory variables to be encoded as numerical. For binary variables, that is just a technical detail to note. For multi-class categorical variables, use one-hot encoding to replace the variable with a collection of binary variables.

Deliverables

Submit your result in pdf format. The deliverable should contain the information specified in points 1 to 4 above.