

UNIVERSITY OF STRASBOURG  
ARTIFICIAL INTELLIGENCE  
Licence 3 Informatique

## Topic 1 - Learning by aggregating independent models

*Fundamental model: decision tree*

### A. Objective

---

For this topic, your objective is to implement and evaluate a classification model built from decision trees. By strategically combining these trees, you will explore how the resulting ensemble can improve the predictive performance of your model.

### B. Work to be done

---

*In this subject, you only need to separate your data once: a training set (80%) and a test set (20%).*

#### 1) Building a decision tree

Using the code developed in the lab, implement a decision tree with a maximum depth of 10, trained on your training set. Your tree will be binary, as in the practical exercises. You will partition the data using the median of the attribute examined as the split value.

#### 2) Data set sampling

1. Implement a function sample  $N$  of the  $N$  instances in the training set, with discount.
  - It is estimated that with such a process, around  $1/3$  of the  $N$  instances will not be sampled. These instances will have to be set aside to form the validation set.
2. Train sets of  $t=1$ ,  $t=4$ , then  $t=8$  trees independently (i.e. within a given set, each tree is trained on a particular sampling:  $t$  trees therefore mean  $t$  distinct samplings).

#### 3) Prediction aggregation

Implement two strategies for aggregating predictions from a set trees:

1. Majority vote: each tree in the set contributes equally. The final prediction is the prediction of the majority of trees in the ensemble.
2. Confidence-weighted voting: a tree's prediction is weighted according to its confidence in the decision taken. Confidence can be expressed as the ratio of classes in the leaf giving rise to the prediction.

#### **4) Model evaluation**

1. Evaluate each of your 3 sets exhaustively on the corresponding validation data.
2. Compare and interpret the results: what differences do you observe between the different sets? How do you think some sets generalize better than others?
3. Make a final evaluation of your best set on the test set.

#### **5) Reflective conclusion**

1. Why does combining different trees deliver better classification performance than a single tree?
2. While models such as decision trees are appreciated for their transparency, since we can, for example, access the most discriminating attribute as calculated by the model, is there any guarantee that the trees in a set all determine the same order of attribute discriminating power?
3. What constraints might make the approach you've implemented impractical or unprofitable?

### **C. Items to be assessed**

---

1. Randomly extract an instance from the test set, make a prediction with a (single) tree and display the prediction, its confidence, and the path taken in the tree.
2. Using a confusion matrix, indicate which class is the most difficult to predict for one of your models.
3. By comparing the training and validation accuracy of one of your models, indicate whether the model has overlearned (explain your answer).
4. Explain why it is consistent to set aside non-sampled instances to form the validation set of an ensemble.
5. Show how set performance changes as the number of trees increases from 1 to 4, then 8.

6. For a few instances of the test set for which you obtain correct and incorrect predictions, show the individual predictions (together with their confidence) of the trees in the set  $t=8$ , and illustrate how the prediction of the set changes according to the aggregation method.
7. Show the most discriminating attribute of each tree in the set  $t=8$ . In what way do you think sampling allows to obtain diversity among the trees in the set?
8. Compare the evaluation metrics of a single tree vs. the  $t=8$  set: explain which metrics have improved most significantly and what this might suggest about the behavior of the set.

## Ongoing Assessment (20%)

The items to be assessed are specific, and it is your responsibility to call your lab instructor to validate each item (i.e., when you are ready, you notify us and we check it).

There will be no partial validations or multiple checks for a given item: at the time of verification, an item is either validated or not.

To avoid every group trying to validate all 8 items during the last lab session, two major milestones are set:

- 4 items must be validated no later than the week of **April 20**
- The remaining 4 items must be validated no later than the week of **May 5**

Anything not validated by the deadline will not be assessed (clear tautology!): if you validate only 3 items by the week of April 28, you will only be allowed to validate 4 out of the remaining 5 items before the second milestone.

The **order of item evaluation** does not matter (though you'll naturally see that some project elements depend on others, so it's unlikely that final-stage items can be validated early on).

---

## Final Assessment (80%)

You must submit a **5–10 page report (everything included)** in **PDF format, no later than the evening of May 19**.

Your report must be properly structured.

The writing should be clear, in an appropriate academic tone, supported by illustrations, graphs, tables, or any other elements that highlight your work. Sources must be cited where relevant.

Your report must be accompanied by an **archive containing your source code**:

- In **.py format**. Notebooks will be accepted **only** for experimental parts. In other words, core models or functions must be implemented in modules, as shown in the decision tree and neural network labs.

Your code must be **clear, well-structured, and documented**.

Your **experimental approach** must be explained well enough to make your tests **reproducible**.