Pattern Recognition
Homework 3 Announcement

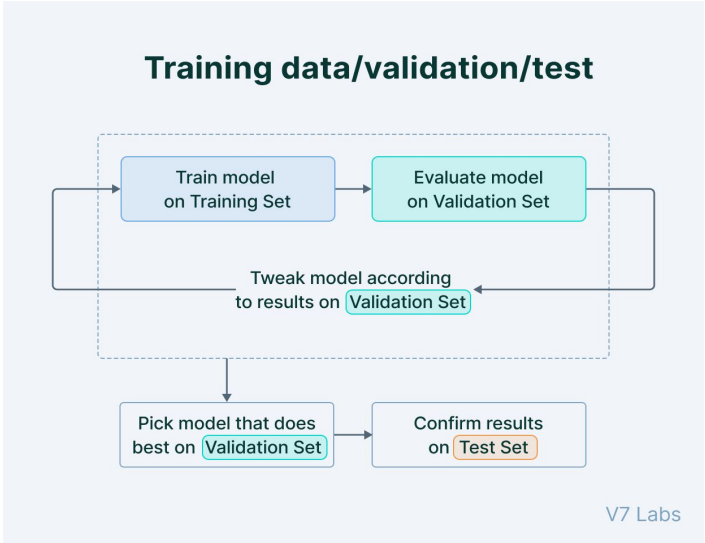
Lastest update: 2023.04.12 12:00

Homework 3

- Deadline: Apr. 26,Wed. at 23:59
  - Code assignment (70%)
    - Implement Decision Tree and Random forest using only NumPy.
  - Questions (30%)
    - Write your answer in detail in the report.
- Question: Link
- Sample code: Link
- Dataset: Link

Dataset

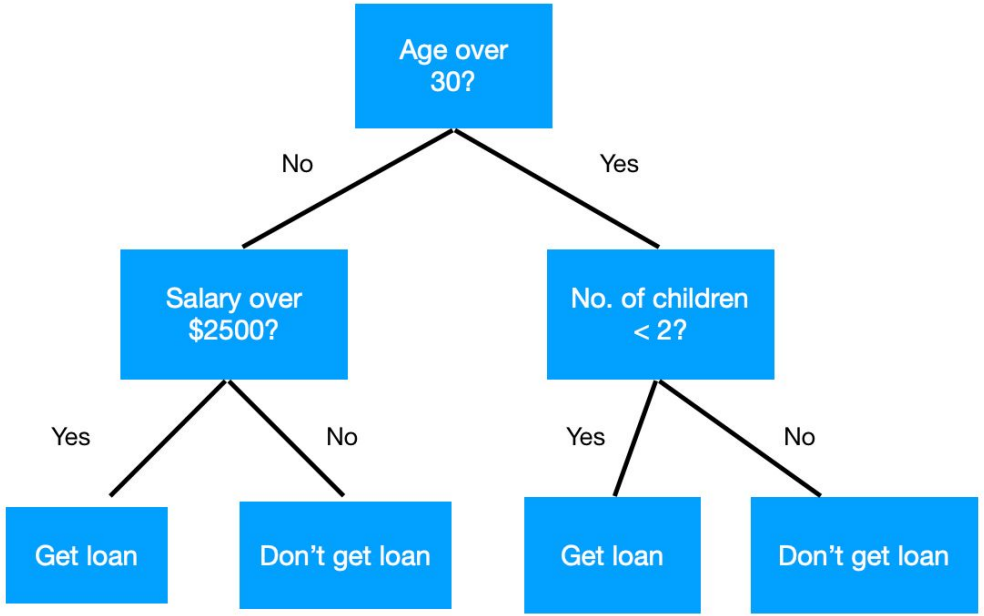
- A real world dataset
  - Training set (800 data)
  - Validation set (800 data)
  - Testing set (800 data)
- 7 features, 7 labels



https://www.v7labs.com/blog/train-validation-test-set

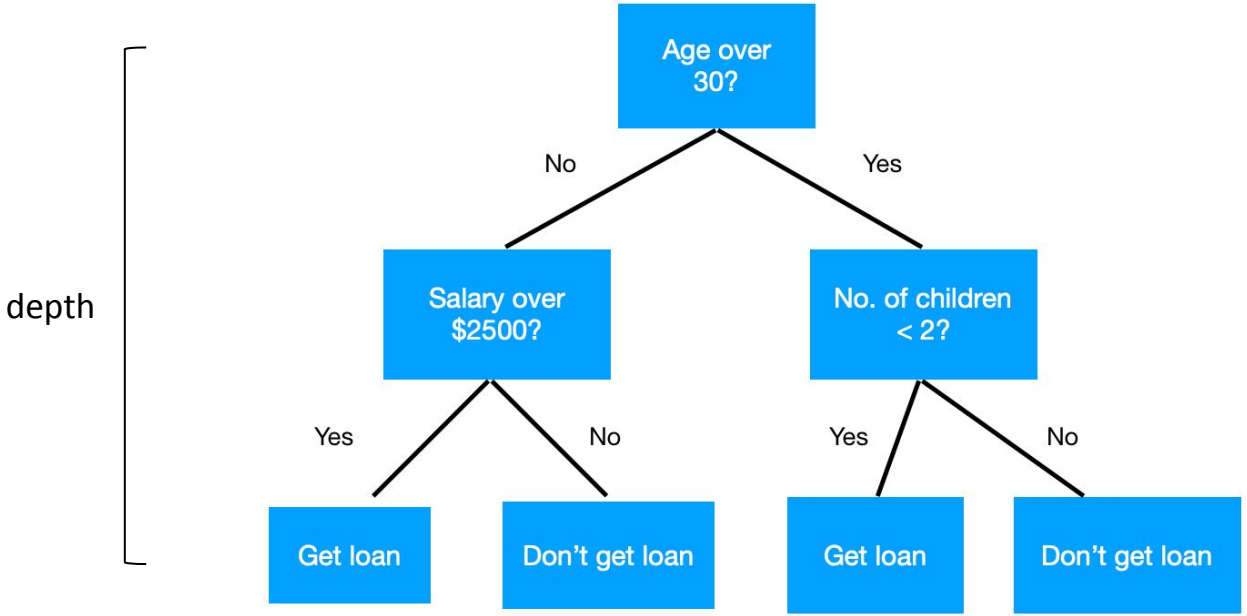
Decision Tree

- The task is to determine whether to approve a loan for a customer.



Decision Tree

- Find the feature that will split the data in a way that results in the most **pure** classes at the resulting nodes.



Decision Tree

- How to measure **“pure”**?
  - Entropy: the smaller, the purer
  - Gini-index: the smaller, the purer

$$Gini = 1 - \sum_j p_j^2$$

	Parent
C0	6
C1	6
Gini = 0.5	

Gini :
$1 - (6/12)^2 - (6/12)^2$
= 0.5

$$Entropy = -\sum_j p_j \log_2 p_j$$

- If all classes are the same in one node

$$entropy = -1 \log_2 1 = 0$$

- If the classes are half-and-half

$$entropy = -0.5 \log_2 0.5 - 0.5 \log_2 0.5 = 1$$

Decision Tree

- Until stopped
  a. Select a node
  b. Loop through all values of all features
     1) Partition the node and calculate the purity of the resulting data
     2) Find the feature value that yields the lowest value of Gini or Entropy
  c. Split the node using the feature value found in step b
  d. Move to the next node and repeat steps a to c
- Stopping criteria
  - The data in each leaf node belongs to the same class
  - The depth of the tree reaches a pre-specified limit

Decision Tree

- Decision tree can find a unique path for each data (over-fitting) if we don’t pre-specified any limits, such as the **depth of the node.**

Questions for Decision Tree (30%)

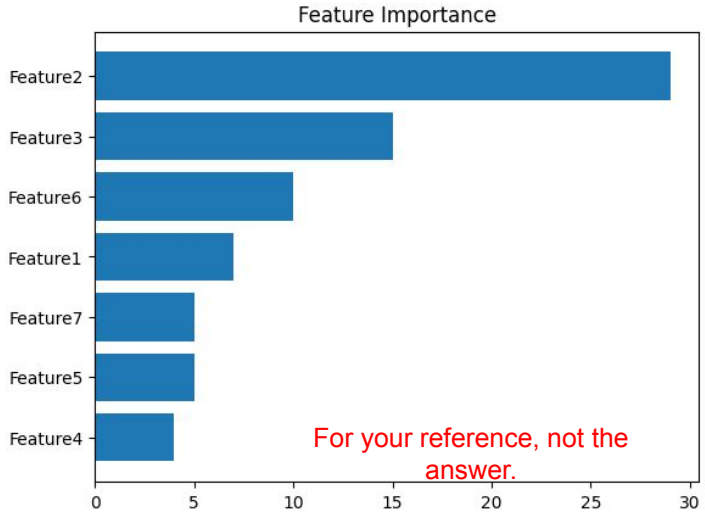
- Q1 (5%): Compute the Entropy and Gini Index of the following inputs.

```
ex1 = np.array(["+", "+", "+", "+", "+", "-"])
ex2 = np.array(["+", "+", "+", "-", "-", "-"])
ex3 = np.array(["+" ,"-", "-", "-", "-", "-"])
```

- Q2 (10%):
  - Fix criterion='gini', max_features=None
  - Try max_depth=3 and max_depth=10
- Q3 (10%):
  - Fix max_depth=3, max_features=None
  - Try criterion='gini' and criterion='entropy'

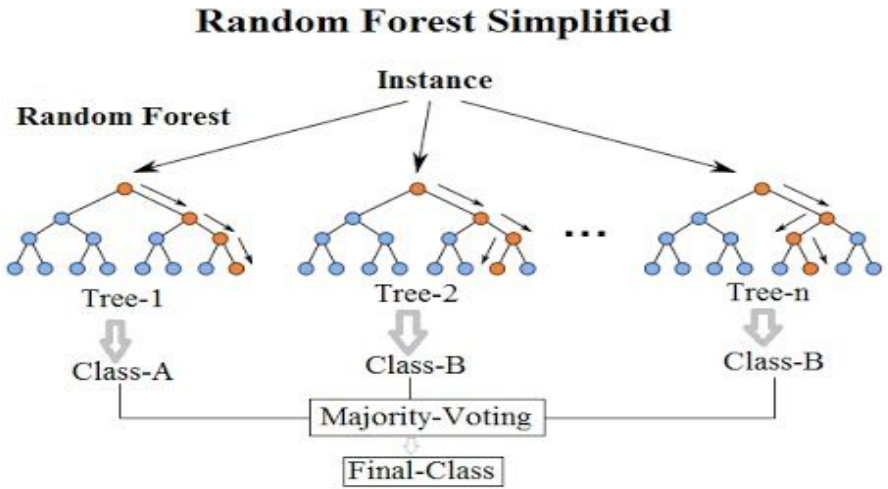
Questions for Decision Tree (30%)

- Q4 (5%):
  - Train your model using criterion='gini', max_depth=10, and max_features=None.
  - Plot the feature importance of your decision tree model by simply counting the number of times each feature is used to split the data.



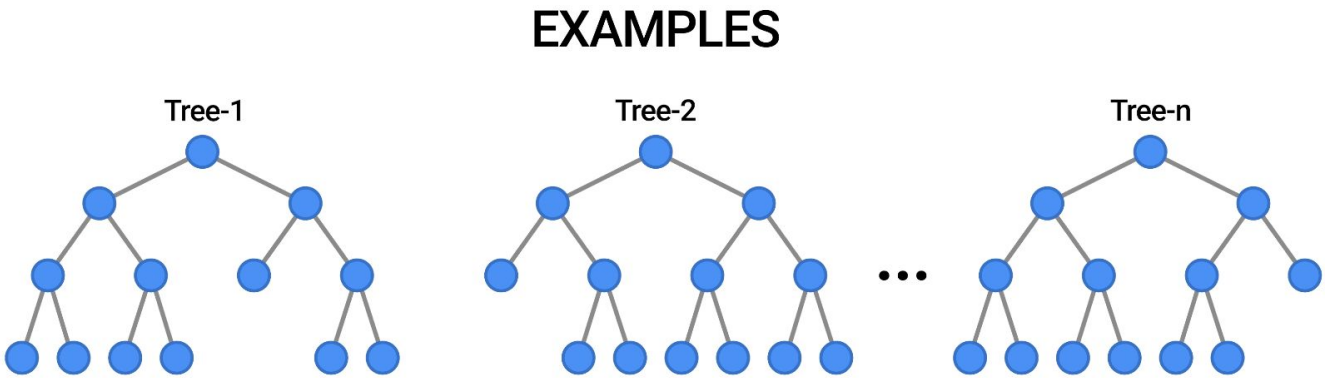
Bagging

- Ensemble method of decision trees.
- Bagging (Bootstrap aggregating): Fit many deep trees to bootstrap-resampled versions of the training data, and classify data by majority voting



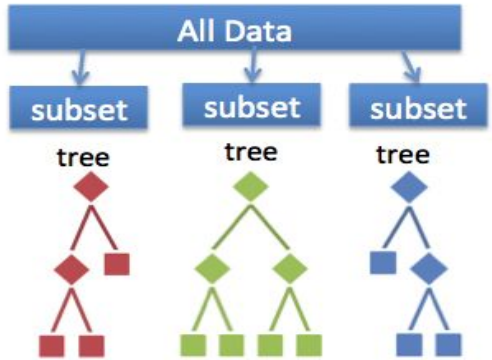
Random Forest

- Random forest combines multiple decision trees to make a better estimator than the individual decision trees.
- Use the Decision Tree you implemented as the weak estimator.



Random Forest

- Where does the **randomness** come from in a random forest algorithm?
  - Bootstrapped datasets
  - Random feature selection during the growth of each tree in the forest.
  - Specifically, each tree may grow using different subsets of the original data and features, which are randomly sampled during the construction process.



Random Forest

Algorithm 1 Random Forest

Precondition: A training set $S := (x_1, y_1), \ldots, (x_n, y_n)$, features F, and number of trees in forest B.

```
1  function RandomForest(S, F)
2      H ← ∅
3      for i ∈ 1, ..., B do
4          S^(i) ← A bootstrap sample from S
5          h_i ← RandomizedTreeLearn(S^(i), F)
6          H ← H ∪ {h_i}
7      end for
8      return H
9  end function
10 function RandomizedTreeLearn(S, F)
11     At each node:
12         f ← very small subset of F
13         Split on best feature in f
14     return The learned tree
15 end function
```

Questions for Random Forest (20%)

- Q5 (10%):
  - Fix criterion='gini', max_features=sqrt(n_features), max_depth=None, and Bootstrap=True
  - Try **n_estimator=10** and **n_estimator=50**.

- Q6 (10%):
  - Fix criterion='gini', max_depth=None, Bootstrap=True, and n_estimator=10
  - Try max_features=sqrt(n_features) and max_features=None.

Train your own model (20%)

- You can use nither Decision Tree or Random Forest that you implemented.
- Try different parameters and features to beat the baseline.
- Explain in detail how you choose the model, parameters, and features in the report. Otherwise, extra penalty.
- Predict the testing data and save the result into a CSV file.

Train your own model (20%)

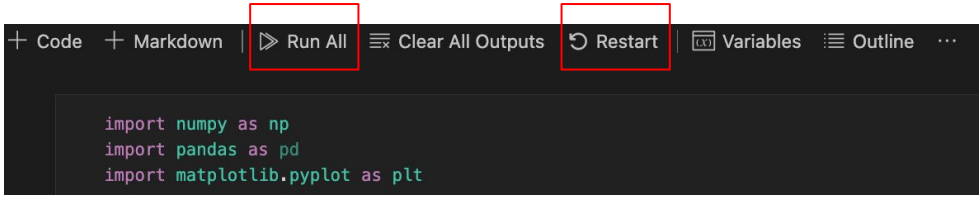
- Evaluation is based on testing accuracy.
- Testing data distribution is guaranteed to be similar to validation data.

Points	Testing Accuracy
20 points	acc > 0.915
15 points	acc > 0.9
10 points	acc > 0.88
5 points	acc > 0.8
0 points	acc <= 0.8

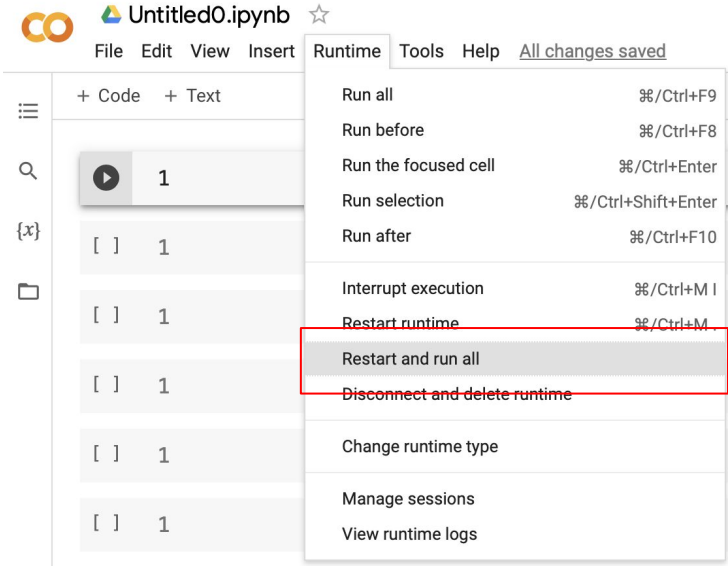
Report

- Please write your report in English.
- Please follow the HW1 report template.
- You must type the answer and also screenshot at the same time for the coding part.
- **Answer each question as clearly as possible**. You will get an extra penalty for only the brief answer.

Submission



- Compress your .ipynb, .pdf, and .csv into a zip file and submit it on E3.
- Before submission:
  - Restart and run All
  - Save and submit the .ipynb (keep all cell outputs)
  - Get 0 points if you do not keep the cell outputs.
- <STUDENT ID>_HW3.zip
  - <STUDENT ID>_HW3.ipynb
  - <STUDENT ID>_HW3.pdf
  - <STUDENT ID>_prediction.csv

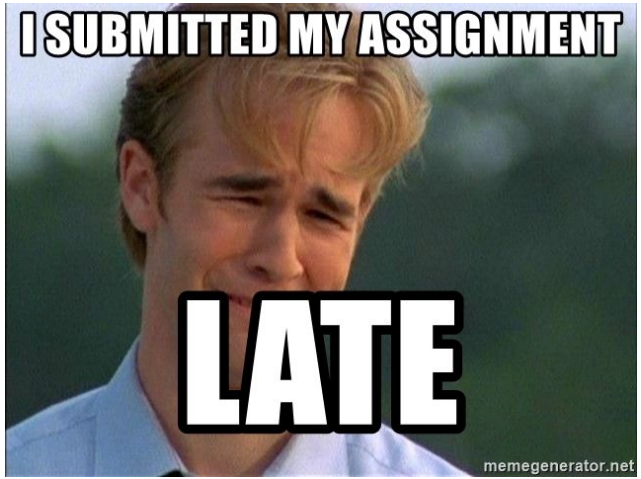



```
❯ zip -r 310551056_HW1.zip 310551056_HW1.ipynb 310551056_HW1.pdf 310551056_prediction.csv
  adding: 310551056_HW1.ipynb (deflated 34%)
  adding: 310551056_HW1.pdf (deflated 8%)
  adding: 310551056_prediction.csv (deflated 57%)
```

For your reference

Late policy

- We will deduct a late penalty of 20 points per additional late day.
- If you get 90 points but delay for two days, you get 90 - (20 x 2) = 50 points!
- We only accept submissions that are up to 10 minutes late. Any submissions that are later than that will be considered late, regardless of the reason.



HW2

- We will announce the HW2 scores before Apr. 16.
- The criterion has been adjusted from the left table to the right table.

Points	Testing Accuracy
20	testing acc > 0.921
15	0.91 < testing acc <= 0.921
8	0.9 < testing acc <= 0.91
0	testing acc <= 0.9

Points	Testing Accuracy
20	testing acc > 0.921
15	testing acc > 0.91
10	testing acc > 0.8
5	testing acc > 0.6
0	testing acc <= 0.6