11310CS460200 Intro 2 ML Lab 2

Name	Date
110020007 施淙綸	2024/10/12

Attributes setting

	# of trees	# of features	# of instances	max # of depth	min # of samples split	percentage per step
#	17	10	$0.98 \cdot 8500 \cdot 0.3 = 2499$	8	500	$\frac{1}{33}$

※Note that 0.98 in `# of instances` is the split ratio of traning data and validation data.

Difficulty I encountered

- **Too many possibilities of split point**: Calculate the information gain of each possibilities is time-consuming.
- **Threshold of gender**: It's trivial to find the best split of gender.

How I solved difficulties and my reflections

- **Too many possibilities of split point**: Using fixed percentiles to calculate the information gain. For further optimization, I write another script using multiprocessing pool to find the best split in each tree building process. It significantly cut down the training time by approximately 60%.
- **Threshold of gender**: I simply return 0.5 as the split point rather than finding the exactly split idx and have identical result.
- **Fine-tuning**: As always, find the best hyperparameters combination is the hardest part of ML training.