Assignment #4

Course: *Machine learning*
Date: *November 10th, 2025*

Assignment

In this assignment, you will learn about decision/regression trees.
You learned about decision/regression trees in your previous courses. You should renew your knowledge by watching: `https://www.youtube.com/watch?v=g9c66TUylZ4`.

Download the dataset "House price". Price is the target. Preprocess it.

The dataset is available at `https://www.kaggle.com/datasets/balakrishcodes/others?select=House_Price.csv`

Implement the regression tree algorithm from scratch.

For the splitting criteria use a criteria of your choice. The residual sum of squares is usually used as the splitting criteria in regression trees. Select a criteria to stop slitting the nodes.
Examples of criteria: select the depth of the tree or select the minimal required number of instances in one leave. This will affect how complex your tree will be.

Build a regression tree for the selected dataset.

Test the regressing tree using cross-validation. Test different stopping criteria.

Observe when your tree starts overfitting. Comment on the results.

Compare the cross-validation results with those you get while building a regression tree with scikit-learn. Use the same cross-validation splits on both models.

Modify your regression tree algorithm to perform as regression trees in random forest.

In random forest trees in each split, only a portion of randomly chosen features is considered.

BONUS (+ 2 points) Run a small random forest (10-50 trees) and get regression result.

Train different RF trees on bootrstrap data and combine the results to get the final result.