Module 13 - Bonus Challenges

§M13: Ensemble Learning

- Below are open-ended bonus challenges; solving them is not required but can help you better understand ML/AI in the context of engineering, and how to use them in practical cases.
- Bonus points earned in all homework assignments will be averaged (6 bonus points for each assignment) and then directly added to your final score to calculate your final letter grade.

Challenge 1.1. In this task, you will compare the performance of a single decision tree classifier with that of a bagging classifier using decision trees as base estimators. The objective is to explore how varying the number of estimators in the Bagging ensemble affects both decision boundaries and model performance. Specifically, you will work with a 2D classification problem generated by the `make_moons` function from the `sklearn.datasets` module, using the following code:

```
X, y = make_moons(n_samples=300, noise=0.2, random_state=42)
```

This dataset consists of 300 samples with a noise level of 0.2.

Using this dataset, your tasks are as follows: (*6pts*)

1. Train a single decision tree.
   - Use 5-fold cross-validation to evaluate the performance. The performance metric should be the accuracy score for each fold.
   - After performing 5-fold cross-validation, fit the model to the entire dataset.
2. Train multiple bagging classifiers using decision trees as the base estimator.
   - Vary the number of estimators as follows: `n_estimators = 1, 10, 20, 50, 100`.
   - Use `bootstrap=True` and set `max_samples=1.0`.

- For each model, use 5-fold cross-validation to evaluate the performance. The performance metric should be the accuracy score for each fold.
- For each model, after performing 5-fold cross-validation, fit the model to the entire dataset.

3. Boundary comparison: For each model trained on the entire dataset, plot the decision boundaries to visualize how the model classifies the dataset. Display these boundaries in a single row of subplots. The first plot should show the decision boundary of the single decision tree, followed by the bagging classifier models with increasing numbers of estimators. Your plots will look like Figure 1.

Figure 1: Example of decision boundary visualization

4. Performance comparison: Create a box plot to compare the accuracy scores from cross-validation. Include the single decision tree and the bagging models with `n_estimators = 1, 10, 20, 50, 100` in the same plot for comparison. Your plot will look like Figure 2.

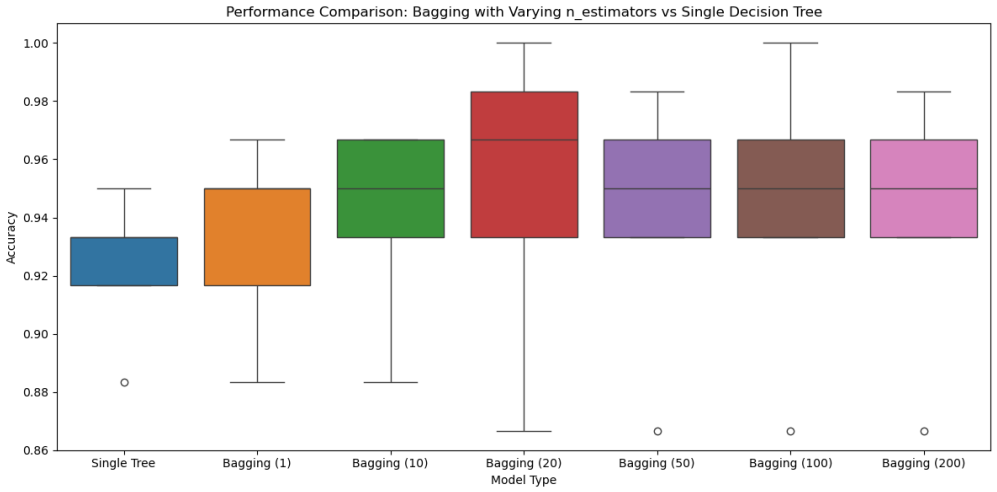

Figure 2: Example of box plot

5. Analyze your results:

- Compare the decision boundaries and performance (accuracy and variance) of the single decision tree with that of the BaggingClassifier.

- Discuss how increasing the number of estimators in the BaggingClassifier affects the decision boundaries and model performance. Is there a point where performance plateaus as the number of estimators increases?

6. Submit your Jupyter Notebook file with appropriate comments.