Module 10 - Bonus Challenges

§M10: Bias-Variance Tradeoff, Model Selection

- 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 bonus question, you will use Scikit-learn's built-in model selector to identify the best model for the UCI Yacht Hydrodynamics dataset[1]. This dataset is stored in the file `'m10_bonus.csv'`, which contains experimental results from various sailing yachts. The objective is to predict the residuary resistance based on parameters related to geometry and flow conditions. These parameters are summarized as follows:

- Input 1: `'LC'` - Longitudinal position of the center of buoyancy
- Input 2: `'PC'` - Prismatic coefficient
- Input 3: `'LD'` - Length-displacement ratio
- Input 4: `'BDr'` - Beam-draught ratio
- Input 5: `'LB'` - Length-beam ratio
- Input 6: `'Fr'` - Froude number
- Output: `'Rr'` - Residuary resistance

We will consider three types of models: KNN, SVM, and Random Forest, each with several hyperparameters to tune, where each hyperparameter has multiple possible choices. The models and their corresponding hyperparameter choices are outlined in Table 1.

Note that exhaustively testing all possible hyperparameter combinations for each model type can lead to a large number of candidates, making manual selection tedious. Fortunately, Scikit-learn provides some built-in functions to automate and expedite this process. Your task is to familiarize yourself with this process and select the best model using the UCI Yacht Hydrodynamics dataset. Please complete the task according to the following instructions: (*6pts*)

Table 1: Models and Their Hyperparameters

Model	Hyperparameter	Possible Choices
KNeighborsRegressor	*model_n_neighbors*	3, 5, 7
SVR	*model_kernel*	linear, rbf
	model_C	0.1, 1, 10
RandomForestRegressor	*model_n_estimators*	100, 200
	model_max_depth	5, 10

1. Learn about the `GridSearchCV` function in `sklearn` for model selection and hyperparameter tuning. You can refer to the following links:
    - GridSearchCV documentation: https://scikit-learn.org/dev/modules/generated/sklearn.model_selection.GridSearchCV.html
    - Example: https://scikit-learn.org/dev/modules/grid_search.html#grid-search
2. Randomly split the dataset into 80% for training and 20% for testing using `train_test_split` from `sklearn`.
3. Use `GridSearchCV` with 5-fold cross-validation to select the best model and its hyperparameters from Table 1. Report the best model and its hyperparameters.
4. Apply the best model to the testing dataset and report the mean squared error.
5. Submit your Jupyter Notebook file with appropriate comments.

References

[1] J. Gerritsma, R. Onnink, and A. Versluis. *Yacht Hydrodynamics*. 2007. URL: https://doi.org/10.24432/C5XG7R.