Exercise 1:

Having mastered the intricacies of random forests, our medical research team has set eyes on yet another (and indeed very central) topic in machine learning: **tuning**. Researcher Lisa, who has vigorously studied the I2ML lecture materials, is asked to explain tuning in simple terms to her colleagues: *"Hyperparameter tuning, often abbreviated as tuning, can be broken down to one simple formula:*

$$\min_{\boldsymbol{\lambda} \in \boldsymbol{\Lambda}} \widehat{\mathrm{GE}}_{\mathcal{D}_{\mathrm{test}}} \left(\mathcal{I}(\mathcal{D}_{\mathrm{train}}, \boldsymbol{\lambda}) \right). \tag{1}$$

Researchers Son and Holger seem to have not studied the materials as engaged as Lisa, since they cannot make sense of the expression above.

1) **Explain hyperparameter tuning in your own words, using the formula above. What do data scientists mean when they call it a bi-level optimization problem?**

A tuning problem consists of several elements:

- A data set $\mathcal{D}$,

- A learner $\mathcal{I}$,

- d hyperparameters of the learner and their configuration space $\boldsymbol{\Lambda} = \Lambda_1 \times \Lambda_2 \times \ldots \times \Lambda_d$,

- A performance measure ρ to estimate the generalization error, as determined by the application

- A (resampling) procedure for estimating the predictive performance

Son finds himself wondering how this would translate to their specific research problem. As before, they try to predict whether a patient admitted to the hospital will require intensive care, a binary classification task with target space $\mathcal{Y} = \{0, 1\}$. The feature space is the same as before: $\mathcal{X} = (\mathbb{R}_0^+)^3$, with $\mathbf{x}^{(i)} = (x_{age}, \, x_{blood \, pressure}, \, x_{weight})^{(i)} \in \mathcal{X}$ for $i = 1, 2, \ldots, n$ observations.

2) **Given a data set $\mathcal{D}$, a random forest learner $\mathcal{I}$, and our research group's problem, write down an example for a specific tuning problem using the list above.**

The research group wants to apply the tuning procedure and get as best an estimate of the generalization error as they can. Holger tunes a learner with a simple 5-fold cross-validation (CV), training 5 models for each of the 100 hyperparameter configurations $\boldsymbol{\lambda}_i$, $i = 1, \ldots, 100$ generated by random search. The unseen CV splits are used to estimate the performance, and the best-performing hyperparameter configuration $\boldsymbol{\lambda}^*$, which yielded an average accuracy of $\rho_{ACC} = 0.92$, is chosen to train the final model on the entire data set.

3) **Does the final model result in an expected accuracy ρ_{ACC} wrt new, unseen observations equal to, lower or higher than 0.92?**