🏁 Wrap-up quiz 2

This quiz requires some programming to be answered.

Open the dataset `blood_transfusion.csv` with the following command:

```
import pandas as pd

blood_transfusion = pd.read_csv("../datasets/blood_transfusion.csv")
target_name = "Class"
data = blood_transfusion.drop(columns=target_name)
target = blood_transfusion[target_name]
```

`blood_transfusion` is a pandas dataframe. The column “Class” contains the target variable.

Question

Select the correct answers from the following proposals.

- a) The problem to be solved is a regression problem
- b) The problem to be solved is a binary classification problem (exactly 2 possible classes)
- c) The problem to be solved is a multiclass classification problem (more than 2 possible classes)
- d) The proportions of the class counts are imbalanced: some classes have more than twice as many rows than others

Select all answers that apply

Hint: `target.unique()`, and `target.value_counts()` are methods that are helpful to answer to this question.

Question

Using a `sklearn.dummy.DummyClassifier` and the strategy `"most_frequent"`, what is the average of the accuracy scores obtained by performing a 10-fold cross-validation?

- a) ~25%
- b) ~50%
- c) ~75%

Select a single answer

Hint: You can check the documentation of `sklearn.model_selection.cross_val_score` here and `sklearn.model_selection.cross_validate` here.

Question

Repeat the previous experiment but compute the balanced accuracy instead of the accuracy score. Pass `scoring="balanced_accuracy"` when calling `cross_validate` or `cross_val_score` functions, the mean score is:

- a) ~25%
- b) ~50%
- c) ~75%

Select a single answer

We will use a [sklearn.neighbors.KNeighborsClassifier] for the remainder of this quiz.

Question

Why is it relevant to add a preprocessing step to scale the data using a `StandardScaler` when working with a `KNeighborsClassifier`?

- a) faster to compute the list of neighbors on scaled data
- b) k-nearest neighbors is based on computing some distances. Features need to be normalized to contribute approximately equally to the distance computation.
- c) This is irrelevant. One could use k-nearest neighbors without normalizing the dataset and get a very similar cross-validation score.

Select a single answer

Create a scikit-learn pipeline (using [sklearn.pipeline.make_pipeline]) where a `StandardScaler` will be used to scale the data followed by a `KNeighborsClassifier`. Use the default hyperparameters.

Question

Inspect the parameters of the created pipeline. What is the value of K, the number of neighbors considered when predicting with the k-nearest neighbors.

- a) 1
- b) 3
- c) 5
- d) 8
- e) 10

Select a single answer

Hint: You can use `model.get_params()` to get the parameters of a scikit-learn estimator.

Question

Set `n_neighbors=1` in the previous model and evaluate it using a 10-fold cross-validation. Use the balanced accuracy as a score. What can you say about this model? Compare the average of the train and test scores to argument your answer.

- a) The model clearly underfits
- b) The model generalizes
- c) The model clearly overfits

Select a single answer

Hint: compute the average test score and the average train score and compare them. Make sure to pass `return_train_score=True` to the `cross_validate` function to also compute the train score.

We will now study the effect of the parameter `n_neighbors` on the train and test score using a validation curve. You can use the following parameter range:

```
param_range = [1, 2, 5, 10, 20, 50, 100, 200, 500]
```

Also, use a 5-fold cross-validation and compute the balanced accuracy score instead of the default accuracy score (check the `scoring` parameter). Finally, plot the average train and test scores for the different value of the hyperparameter. We recall that the name of the parameter can be found using `model.get_params()`.

Question

Select the true affirmations stated below:

- a) The model underfits for a range of `n_neighbors` values between 1 to 10
- b) The model underfits for a range of `n_neighbors` values between 10 to 100
- c) The model underfits for a range of `n_neighbors` values between 100 to 500

Select a single answer

Question

Select the true affirmations stated below:

- a) The model overfits for a range of `n_neighbors` values between 1 to 10
- b) The model overfits for a range of `n_neighbors` values between 10 to 100
- c) The model overfits for a range of `n_neighbors` values between 100 to 500

Select a single answer

Question

Select which of the following statements are true:

- a) The model best generalizes for a range of `n_neighbors` values between 1 to 10
- b) The model best generalizes for a range of `n_neighbors` values between 10 to 100
- c) The model best generalizes for a range of `n_neighbors` values between 100 to 500

Select a single answer