- Welcome
- Introduction: Machine Learning concepts
- Module 1. The Predictive Modeling Pipeline
- **Module 2. Selecting the best model**
  - **Module overview**
  - **Overfitting and Underfitting**
    Quiz M2
  - **Validation and learning curves**
    Quiz M2
  - **Bias versus variance trade-off**
    Quiz M2
  - **Wrap-up quiz**
    Wrap-up quiz
  - **Main Take-away**
- Module 3. Hyperparameter tuning
- Module 4.

Wrap-up quiz 2

In this wrap-up quiz you will need to write some code in order to answer quiz questions:

- an empty notebook is available just below to write your code
- quiz questions are located after the notebook here
- the button Open Notebook at the bottom right of the screen allows you to open the notebook in full page at any time

+ Click here to see a demo video of the notebook user interface

Open Notebook

- Module 5. Decision tree models
- Module 6. Ensemble of models
- Module 7. Evaluating model performance
- Conclusion
- Appendix

Wrap-Up Quiz 2

Importing Pandas

In [2]:
```
import pandas as pd
```

Loading Data

In [11]:
```
blood_transfusion = pd.read_csv("../datasets/blood
target_name = "Class"
data = blood_transfusion.drop(columns=target_name)
target = blood_transfusion[target_name]
# Checking data
data.head()
```

Out[11]:

	Recency	Frequency	Monetary	Time
0	2	50	12500	98
1	0	13	3250	28
2	1	16	4000	35
3	2	20	5000	45
4	1	24	6000	77

Checking target type and imbalance

In [6]:
```
target.value_counts()
```

Out[6]:
```
Class
not donated    570
donated        178
Name: count, dtype: int64
```


Powered by OVHcloud

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

```
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 1 (1/1 point)

Select the correct answers from the following proposals.

- [ ] a) The problem to be solved is a regression problem
- [x] 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)
- [x] 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.

You have used 1 of 2 submissions

Question 2 (1/1 point)

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?

- [ ] b) ~50%
- [x] c) ~75%

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

You have used 1 of 1 submissions

Question 3 (1/1 point)

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%
- [x] b) ~50%
- [ ] c) ~75%

You have used 1 of 1 submissions

Question 4 (1/1 point)

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

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

- [x] 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.

You have used 1 of 1 submissions

Question 5 (1/1 point)

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.

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

- [x] c) 5

- [ ] d) 8

- [ ] e) 10

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

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

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.

You have used 1 of 1 submissions

Question 7 (1/1 point)

We 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:

```
import numpy as np
param_range = np.array([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()`.

Select the true affirmations stated below:

- ○ 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

You have used 1 of 1 submissions

Question 8 (1/1 point)

Select the most correct of the 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

You have used 1 of 1 submissions

Question 9 (1/1 point)

Select the most correct of the affirmations stated below:

- ○ a) The model best generalizes for a range of `n_neighbors` values between 1 to 10

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

You have used 1 of 1 submissions

YOUR EXPERIENCE

According to you, the 'Wrap-up Quiz' of this module was:

- **Too easy, I got bored**
- **Adapted to my skills**
- **Difficult but I was able to follow**
- **Too difficult**

Submit

To answer this wrap-up quiz, I spent:

- **less than 30 minutes**
- **30 min to 1 hour**
- **1 to 2 hours**
- **2 to 4 hours**
- **more than 4 hours**
- **I don't know**

Submit

FORUM (EXTERNAL RESOURCE)

New Topic

Topic	Replies	Last reply
No problem - just saying thanks JCForszpaniak 6d	1	5d ArturoAmorQ
Please add validation rule to answers AlexChan 2 Dec	1	2 Dec glemaitre
Question 6 plinglin 19 Nov	4	20 Nov plinglin

There are no more M2. Wrap-up quiz 2 topics. Ready to start a new conversation?

Terms and conditions