Université de Strasbourg

Practical: Introduction to Machine-Learning
Briefing & Correction

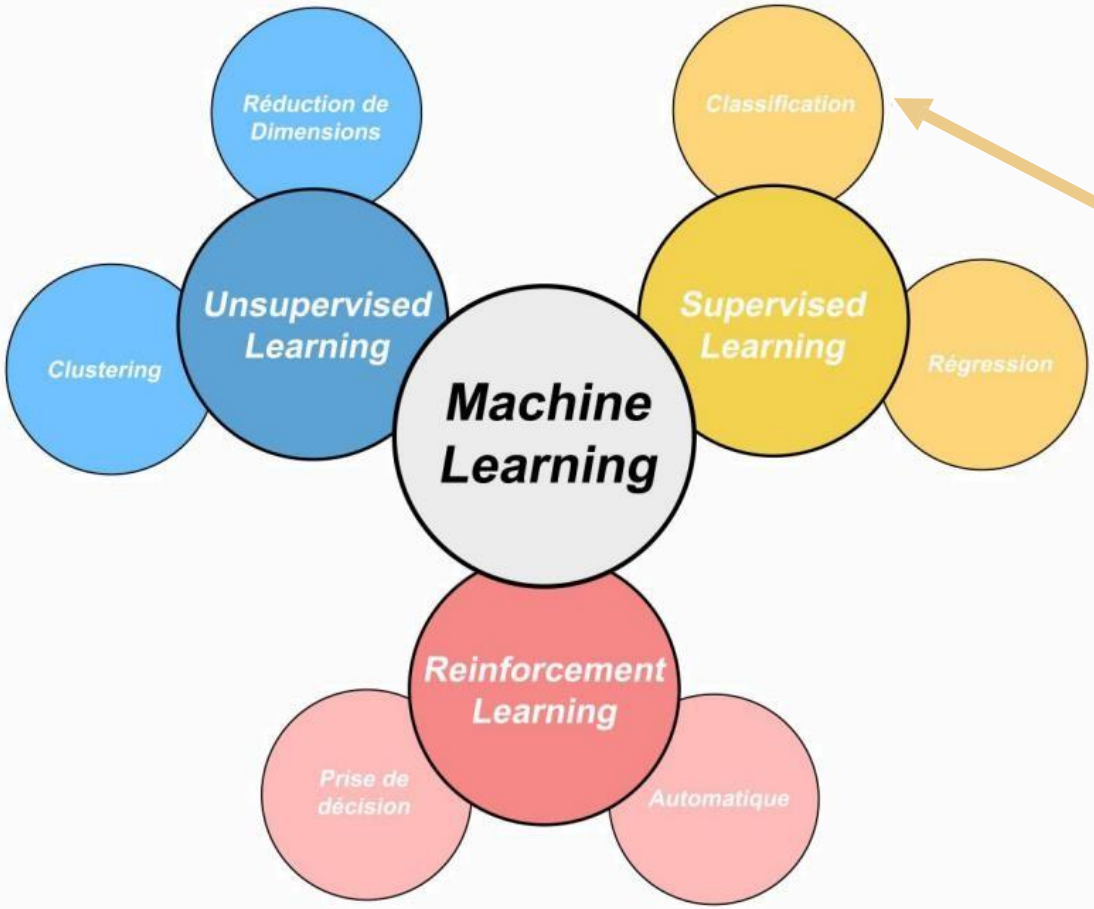
17/10/2023

Nicolas Haas

2nd Year PhD Student
Team CSTB – ICube – CNRS – University of Strasbourg

Contibutors: Created by Corentin Meyer
Adapted by Nicolas Haas

What kind of machine-learning task we will do



Machine-learning: automatically learn from data and make predictions

Today's work: Supervised-Learning -> Classification ->Binary Classification

(Classification: predict a category
Regression: predict a value)

Getting Started

Coding envrionnement
Google Colab online
Jupyter Notebook (Python)

https://colab.research.google.com/

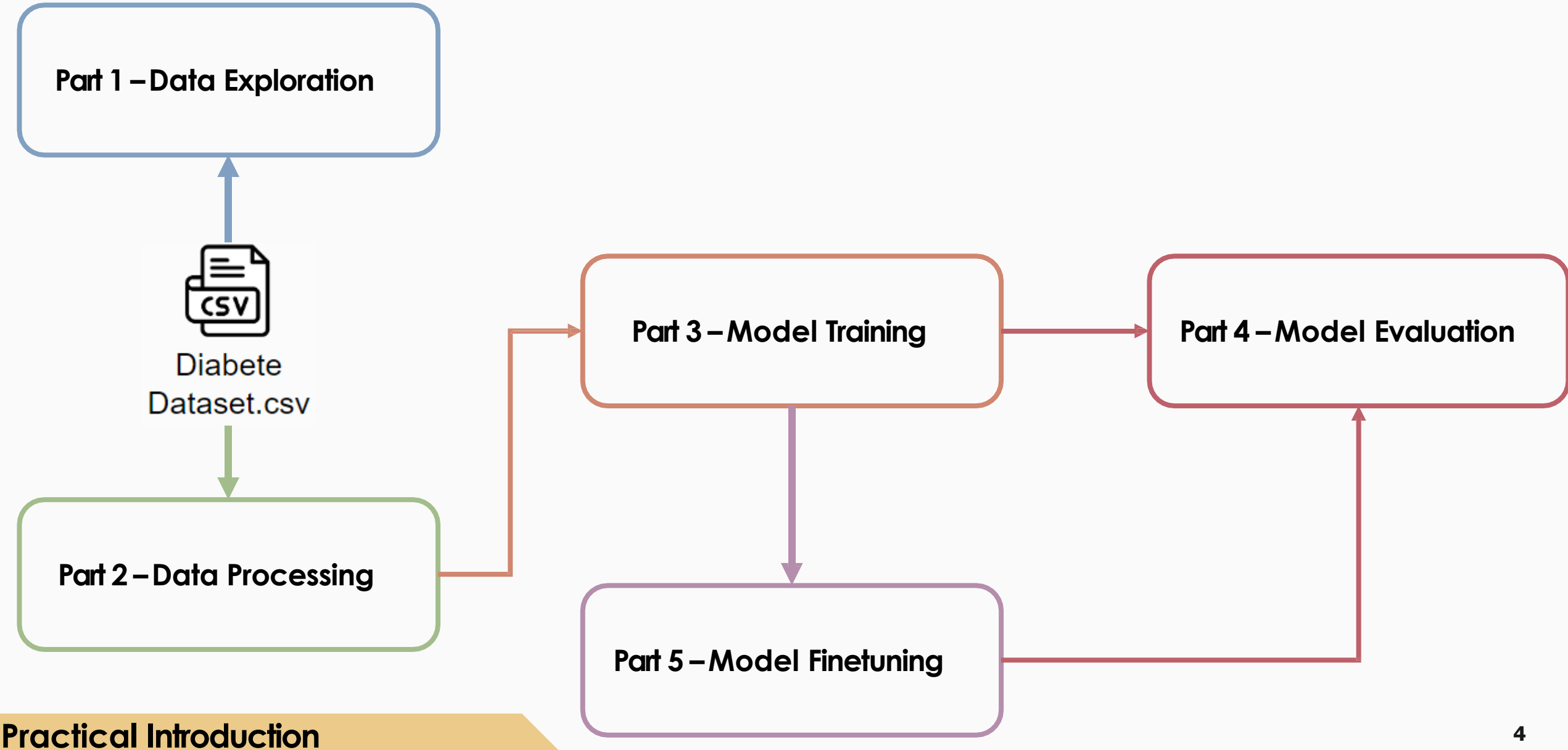
Dataset we will use

Diabetes Dataset

~5.500 persons
17 measurements per person: age, BMI, physical health, mental health…
To **predict the diabetes risk**

Access the practical on GitHub: https://github.com/Dichopsis/ML-TP-ESBS

The five parts of this practical



Part 1 – Data Exploration

Task: do some basic exploratory data analysis

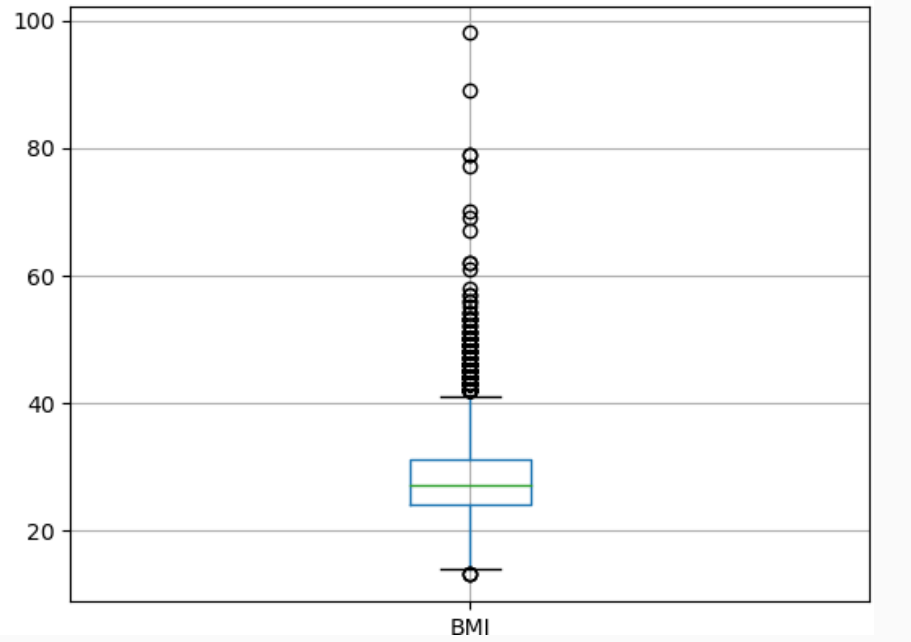
Questions:

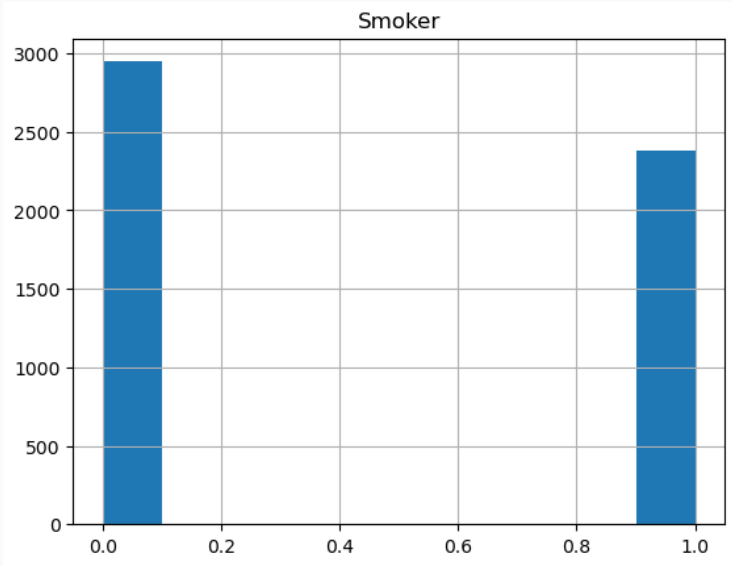
1. How many persons are in this dataset ? (rows). How many features/measurements ? (columns)
5631 rows and 18 columns (17 features, 1 label)

2. What is the percentage of persons with diabetes ?
8.2% with diabetes (label 1), 91.8% without

3. What is the median BMI of the persons in the dataset ? (approximately with BoxPlot)
The median BMI is around 25

4. Is there more non-smoker or smokers in the dataset ? (Histogram)
Around 3000 non-smoker and 2500 smoker.





Part 2 – Data Processing

Task: encode the data to be usable for training by a ML algorithm

Questions:

1. What columns are categorical data, what columns are numeric.

Most columns are categorical except: Age, BMI, Mental Health, Physical Health.

2. What columns are already ready to be used and needs no change.

Most of the columns are usable except for some categorical (sex, highchol, cholcheck, genHlth, MentHtlh, PhysHtlh) and some numerical (Age BMI)

3. What type of processing do you need to do on categorical data and why

Ordinal Encoding ! (0, 1, 2… to n-1 classes)

4. What type of processing do you need to do on numeric data and why

Scaling between 0, +1 or -1, +1

5. What columns contains missing data ? What type of processing do you need to do in this case.

BMI, Smoker and Stroke columns are missing some value (not 5631 non-null content)

Part 2 – Data Processing

Task: Split our dataset between train and test set

Questions:

1. What train/test ratio should you use. A value between 20 and 40% is good

2. How many entries are in your train dataset and in your test dataset.

For 25% I get 4223 persons in train set and 1408 in test set.

3. Verify that you have the same diabetes / diabetes ratio between train and test dataset.
I get 9% of diabetes in training and 8% in testing so it's fine !

Part 3 – Model Training

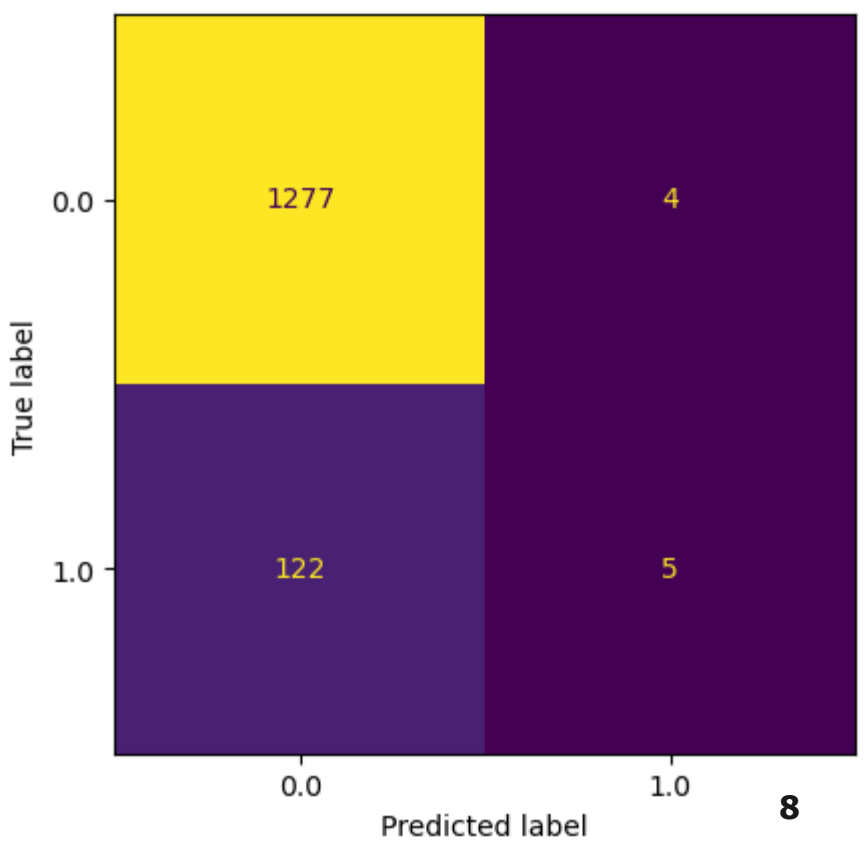
Task: Choose a model a do basic evaluation

Questions:

1. Which model did you choose and why ? Have you set any particular (hyper)parameters ?
Ichoose a RandomForest because well known and used in biology. I also used the class_weight="balanced" to try to go against the imbalanced dataset.

2. What accuracy-score do you get and what conclusion can you take ?
0.91% it's looking pretty good !

3. What do you observe on the confusion matrix and what conclusion can you take ?
Almost all testing data has been classified as "non diabetes", the 91% accuracy is actually misleading !



Part 3 – Model Training

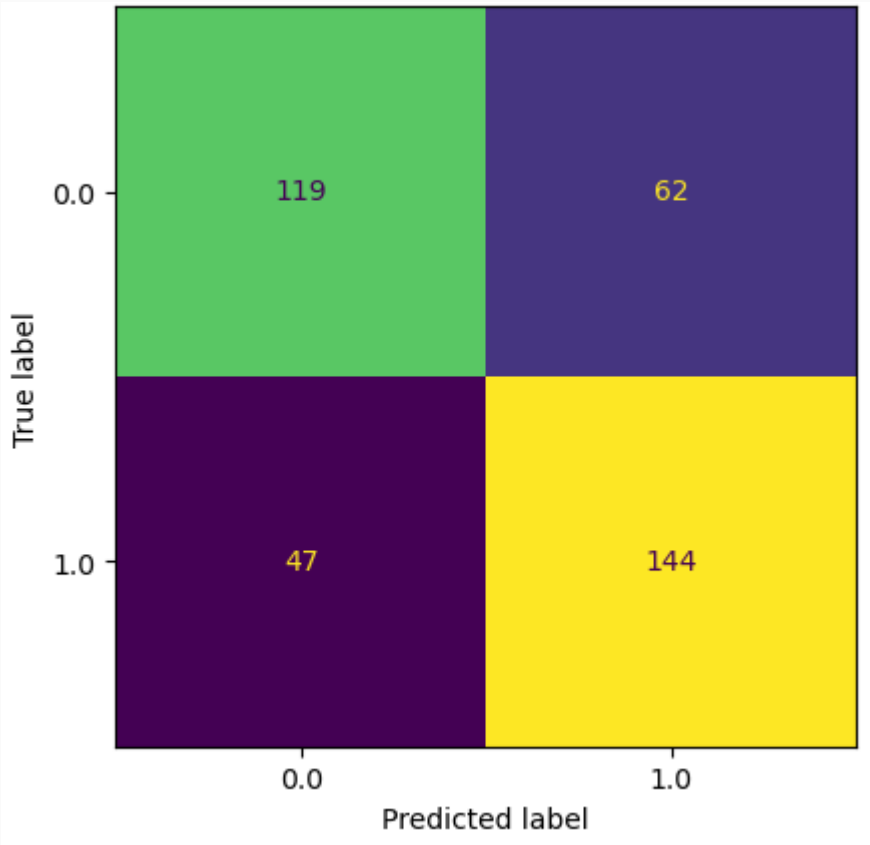
Task: Correcting the previous issue

Questions:

1, What accuracy-score do you get with the new model and what conclusion can you take.
70% accuracy, looks worse, but is it ?

2. What do you observe on the confusion matrix and what conclusion can you take.
At least now our model can predict diabetes in part of the people

3. Did you managed to print the probability of each prediction ? What's the shape of the prediction probability output ? Is there a high variance between the different test entries in probability ?
Shape: n entries x 2 columns. Some datapoints are 50/50% probability, some are 95%/5%.



Part 4 – Model Evaluation

	Balanced-Accuracy	Accuracy	F1-Score	Sensitivity (Recall)	Specificity	Precision	TP	TN	FP	FN
CLF	0.518124	0.910511	0.073529	0.039370	0.996877	0.555556	5	1277	4	122
CLF DownSampled	0.705693	0.706989	0.725441	0.753927	0.657459	0.699029	144	119	62	47

Task: See all common metrics and evaluate both models

Questions:

1. Which model have the accuracy ?

The first one (non downsampled) has the best accuracy.

2. Do you know the difference between accuracy and balanced accuracy ? What model have the best balanced-accuracy

The balanced accuracy is the mean of the accuracy for each class. The second model have a better balanced-accuracy.

3. Which model have the best F1-Score and sensitivity ? Do you know how is F1-Score calculated ?

The second model has the best F1-Score and sensitivity. The F1 Score is the mean between precision and recall (sensibility)

4. Eventually, which model is better according to you based on the metrics ?

From all metrics except the basic accuracy, the second model is clearly the winner. (Look at F1 Score !)

Part 5 – Model Finetuning

Task: Fine tune our model to improve its performances

Questions:

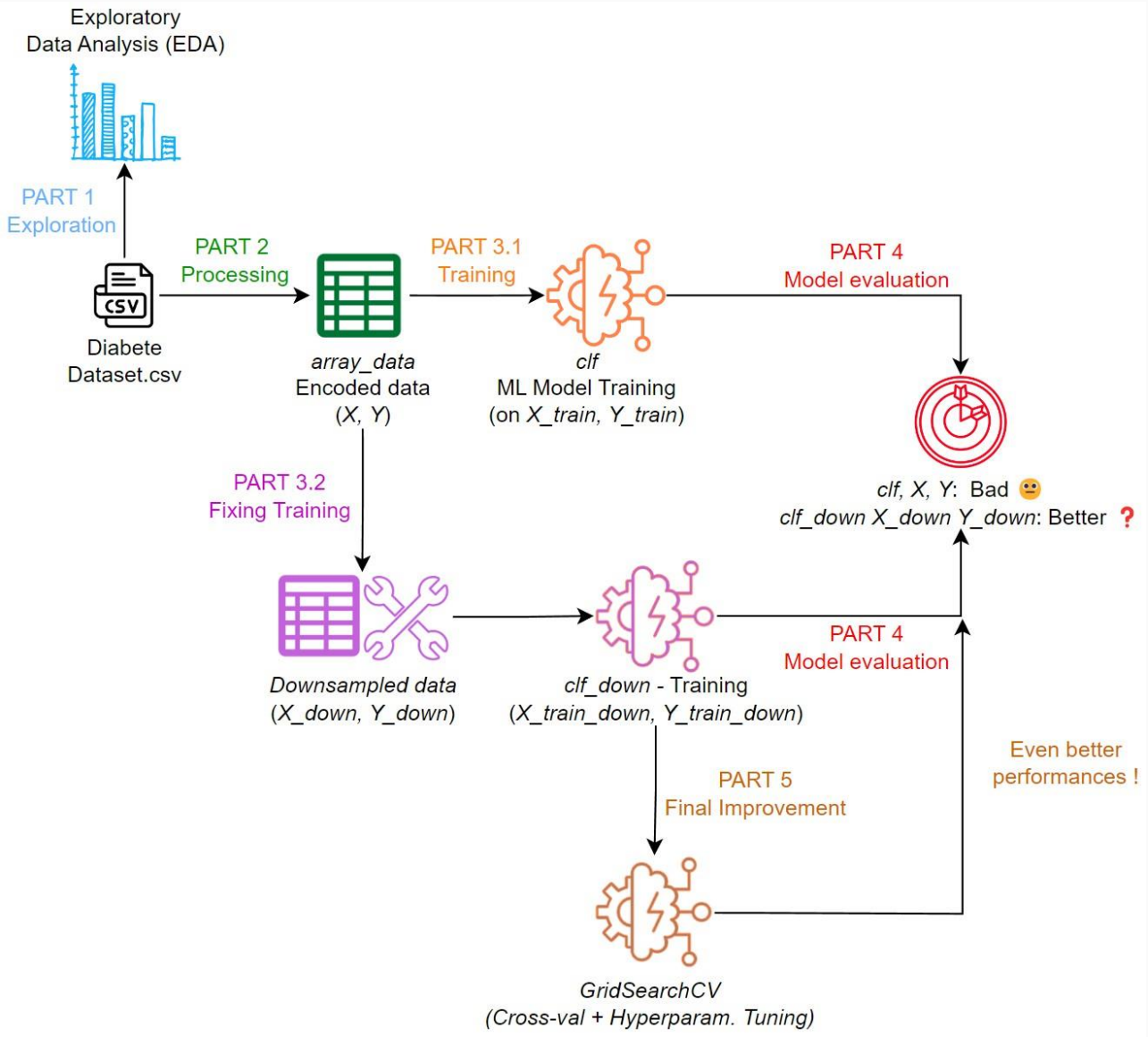
1. What is the point of cross-validation ? Does it increase performance ? If not, what is it useful for ? What is the maximum of fold you theoretically do for a cross-validation ?
Cross-validation doesn't improve performance, it calculate X models on X different test/train splits. It gives a confidence interval on your performances metrics so they are more robust !

2. For GridSearch, you need to choose what metric you want to maximize, what would you choose ?
Balanced metrics or F1-Score is a good choice.

3. What are the parameters for the try with the best metric ? For my own param grids for my Random Forest I get:
Best Parameters: {'criterion': 'log_loss', 'max_depth': 5, 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 50}

4. Compare the metric of the best model found using GridSearch to the metrics of the previous clf_down, is it better now ?
Old balanced-accuracy: 0.70, New: 0.75, so we have an improvement !

Conclusions



You made it !

In this practical you:

- Explored a dataset
- Prepare data for ML
- Trained a ML model and fixed issues
- Evaluate all important metrics of your model
- Optimized your model performance !

Don't hesitate to check Moodle to find further resources on ML if you like this topic. You can also contact me at ni.haas@unistra.fr