# 11752 Machine Learning
# Master in Intelligent Systems
# Universitat de les Illes Balears

## Handout #6: **Supervised learning**
## (graded assignment)

This handout deals with datasets coming from sensor readings on people asked to perform one set of 10 repetitions of the *Unilateral Dumbbell Biceps Curl*. The exercises were performed in different sessions (and maybe on different dates) by six male participants with little weight lifting experience aged between $20 - 28$ years. Participants were informed about how to perform the exercise correctly by an experienced weight lifter. Later, she supervised the execution of all exercises and classified them in five different categories: performed exactly according to the specification (class A), throwing the elbows to the front (class B), lifting the dumbbell only halfway (class C), lowering the dumbbell only halfway (class D) and throwing the hips to the front (class E). Clearly, class A corresponds to the correct execution of the exercise, while the other four classes correspond to common mistakes.

Each group is expected to load and work with the dataset corresponding to the group number. In the course web page, you can find the list of group numbers and the datasets, with filenames `ds_xx.csv` (`xx` is the group number). Each dataset comprises a total of 160 columns involving identification and timestamping data, together with the aforementioned sensor readings. Those readings come from *Inertial Measurement Units* (IMU) able to provide 13 measurements over 1-second time windows:

1. average angular rates for axes x, y and z;
2. average acceleration for axes x, y, z, and the mean total acceleration;
3. average magnetic field for axes x, y and z;
4. average orientation of the Euler angles roll, pitch and yaw;

and several other statistical calculations, like maximum and minimum values, variance, standard deviation, kurtosis and skewness. The devices were attached to the arm and the forearm of the participant, as well as on its belt and the dumbbell involved in the exercise.

---

The tasks to be implemented are described next:

- **You are supposed to design a multi-class classifier** using the data-preparation and the supervised classifier-design techniques considered in the classroom. The proposed classifier is expected to **achieve an $f_1$-score above 90%**.
- Regarding the **preparation of the dataset** you can remove identification data, consider data normalization strategies, use outlier detection algorithms to discard outliers and get rid of features with insufficient data.
- Besides, to shorten computation times, **you are expected to reduce the dimensionality** of the samples but at the same time keep enough features so as to account for the 99% of the variance in the data.
- Regarding **classifier design** you are asked to consider both single-classifier strategies (at least two) and classifier ensembles (at least two).

You are expected to deliver a **jupyter notebook** with the following structure:

1. A brief description of the data preparation strategy adopted.
2. Source code for data preparation.
3. A brief description of the one-classifier attempts you have performed: classification model considered, values of the hyperparameters employed, whether you have used grid search for tuning and the combinations considered, and the performance achieved in terms of $f_1$-score for each.

---

4. Source code for the best strategy based on a single classifier (may depend on having run the data preparation source code first). After cross-validation, provide the average and standard deviation of the scores returned by cross-validation, as well as the confusion matrix for the full dataset and the underlying accuracy.
5. A brief description that enumerates and describes briefly the classifier ensembles implemented, and the performance achieved in terms of $f_1$-score for each.
6. Source code for the best classifier ensemble (may depend on having run the data preparation source code first). After cross-validation, provide the average and standard deviation of the scores returned by cross-validation, as well as the confusion matrix for the full dataset and the underlying accuracy.
7. Source code for processing the common samples included in the `common.csv` file. Provide the confusion matrix and the underlying accuracy.

At least one of the two classifiers at 4. and 6. is supposed to attain an $f_1$-score above 90%.

---

- Delivery date: December 22, 2025. You have to deliver the notebook (.ipynb file) and the corresponding PDF file inside a ZIP file.
- Provide the results requested and suitable comments in the source code.
- This work can be done in groups of two people or individually (see the number of group in the course web page).
- IMPORTANT NOTICE: An excessive similarity between the source code/report released can be considered a kind of plagiarism.