The dataset to be analysed includes measurements from scanned images of fine-needle aspirate of a breast mass. The values represent characteristics of the cell nuclei present in the digital image[1].

The breast cancer data includes 569 examples of cancer biopsies, each with 32 features. One feature is an identification number, another is the cancer diagnosis, and 30 are numeric-valued laboratory measurements. The diagnosis is coded as M to indicate malignant cell or B to indicate benign cell. The 30 numeric measurements comprise the mean, standard error, and worst (that is, largest) value for 10 different characteristics of the digitized cell nuclei, these include: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension.

1. Start by loading the data set ”wisc_bc_data.csv”.
2. Do the graphical exploration of mean attributes in function of the objective attribute: diagnosis, through boxplots and histograms.

3. Develop models for diagnosis prediction using the following algorithms:
   i. *Naive Bayes*
   ii. *k-Nearest Neighbors*

4. Optimize the parameters of the previous algorithms using the **grid search** strategy.

5. Using the sklearn's VotingClassifier class make the **voting combination** of the optimized models by:
   i. Majority vote
   ii. Weighted majority vote

6. Make a meta-learning combination (**stacking**) of the optimized models using the algorithms:
   i. LogisticRegression
   ii. Support Vector Machines

[1] “Breast Cancer Wisconsin (Diagnostic) Data Set” from the UCI Machine Learning Repository

7. Develop diagnosis prediction models by sampling

Bagging:

i. BaggingClassifier

ii. Random forest

Boosting:

iii. Adaboost

iv. Gradient Boosting

v. XGBoost2

vi. LightGBM

8. Compare the performance of different algorithms.

[2] XGBoost Installation Guide: https://xgboost.readthedocs.io/en/stable/install.html#conda