Lab work 1: Binary Classification

1. Goal

Study and assess the performance of SVM classifier used for a decision-making application. You will be using a kaggle data source, the so called “Drinking_Water_Potability” [1], a public usable data source. It provides more than 1000 samples with 9 attributes and its “Y” which designates the potability of the water.
So, the goal is to **tune the binary classifier** such that the optimal value of the obtained **accuracy** using the test data for the potability “Y” is achieved.
Potability = 1 means water is Potable and Potability = 0 means water is not potable.

[1] https://www.kaggle.com/artimule/drinking-water-probability/version/1

2. After completion you have learned

- The importance of Histogram and density analysis
- Applying cross validation
- Applying the “gridsearch” to find the optimal Hyperparameter for the best Accuracy.
- Python and using SVM employing the scikit learn libs

3. Tasks

A group consist of up to two students.
If a single person builds a group, assignment e) and f) are not required.

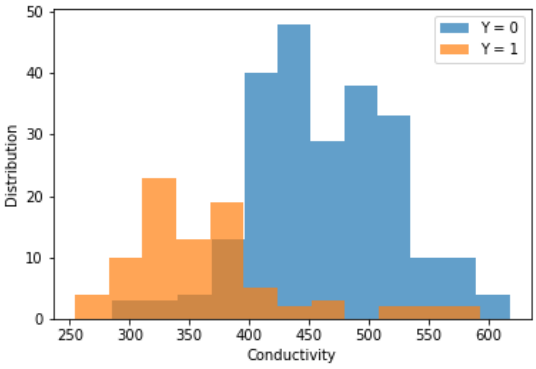
a) Display the Histogram of all attributes including Y

b) Display histogram of each attribute regarding Y=0 and Y=1
What is your conclusion regarding the **expected performance** of the classifier?
Also explain why you came to this conclusion.

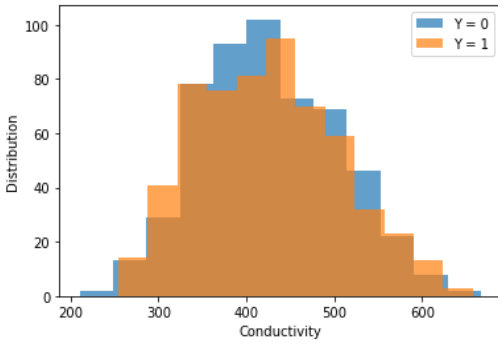
c) Perform serval runs employing the SVM with different C and gamma parameter using **cross validation**. (no shuffle). Vary **n_splits** parameter between 10 and 50.
Display the confusion matrix of the test data for each run.
Calculate the obtained **average Accuracy** of the test data after 5 runs.

d) Perform 5 runs employing SVM with the **gridsearch**
Display best hyperparameter results obtained for each run.
Display the confusion matrix of the test data for each run.
Calculate the obtained **average Accuracy** of the test data after 5 runs.

e) Improve the accuracy by sampling/deleting samples of one attribute e.g., “conductivity” of the source file such that a clear distinction of the histogram for Y=0 and Y=1 is visible, see example below.



Possible histogram after deleting Conductivity attribute samples of the source file



Histogram of Conductivity attribute with the entire source file

f) Display histogram and of each attribute and Perform 5 runs with the **modified data** source employing SVM with the gridsearch.
Display best hyperparameter results obtained for each run.
Display the confusion matrix of the test data for each run.
Calculate the obtained average Accuracy of the test data after 5 runs.
Comment your results.

4. **Submission/presentation**
   Each group submits a **small report (pdf) via email** to the lecturer **3 days before presentation**, containing:

a) source file
b) Histogram of original source and the modified (group of two only), i.e., after deleting source samples of one attribute
c) Confusion matrix and accuracy obtained after employing SVM with cross validation.
d) Confusion matrix and accuracy obtained after employing SVM gridsearch.
e) Confusion matrix and accuracy obtained with modified source after employing SVM gridsearch (group of two only).
f) Conclusion statement
g) The running code is presented and explained by the group.

5. **Remarks**
   Use the source file: drinking_water_potability-1100.csv
   Use a train/test split of 80/20
   Chose for the SVM Hyperparameter, e.g.:
   grid_param = {'C': [1, 10, 100], 'gamma': [1, 0.1, 0.01],'kernel': ['linear','rbf']}
   Use the proper classes of the scikit learn libs

6. **Literature**
   https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html
   https://scikit-learn.org/stable/modules/grid_search.html
   https://scikit-learn.org/stable/auto_examples/svm/plot_rbf_parameters.html
   https://scikit-learn.org/stable/modules/svm.html
   https://scikit-learn.org/stable/auto_examples/svm/plot_svm_kernels.html