Lab work 2: Multi Class Classification

1. **Goal**

   Study and assess the performance of the two different multi class classification approaches discussed in the lecture. You will be using a kaggle data source, the so called "wine quality" [1], a public usable data source. It provides more than 1000 samples with 11 attributes and its "Y" which designates the quality of a red wine.

   The original Kaggle data source was modified towards having now three targets Y, i.e., the wine quality or classes:
   Y=0 poor quality for {Y=2,3,4},
   Y=1 medium quality for {Y=5,6}
   Y=2 premium quality for {Y=7,8}

[1] https://www.kaggle.com/datasets/yasserh/wine-quality-dataset

2. **After completion you have learned**

- The importance of balanced data sources
- Applying the multi class classification using scikit learn libs
- Display further assessment parameter apart from the Accuracy

3. **Tasks**

   A group consist of up to two students.
   If a single person builds a group, assignment d) is not required.

a) Display the histogram of all attributes including Y
   What can you conclude regarding the distribution of Y {0,1,2} ?

b) Display histogram of each attribute regarding Y=0, Y=1 and Y=2
   What is your conclusion regarding the **expected performance** of the classifier?

c) Calculate and compare the obtained test accuracy, precision and recall **with the entire data** source by using:
   1. **One versus All Classifier**
   2. **One versus One Classifier**

   Use estimator=GaussianProcessClassifier().
   You do not need to apply CV this time, one run only is sufficient.
   Optional: Plot the obtained ROC curves and comment it.

d) Improve the accuracy by **deleting specific samples** of the source data (only group of two). Calculate and compare the obtained test accuracy, precision and recall **with the reduced data** source by using:
   1. **One versus All Classifier**
   2. **One versus One Classifier.**

   Use estimator=GaussianProcessClassifier().
   You do not need to apply CV this time, one run only is sufficient.
   Optional: Plot the obtained ROC curves and comment it.

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) Test confusion matrix and test accuracy, precision and recall with original data source for both
   - **One versus All Classifier**
   - **One versus One Classifier.**
   - Comment results
d) Test confusion matrix and test accuracy, precision and recall with modified (deleted samples) data source for both (only group of two)
   - **One versus All Classifier**
   - **One versus One Classifier.**
   - Comment results

e) The running code is presented and explained by the group.

5. **Remarks**
   Use the Kaggle source file: "Wine_Test_02.csv"
   Use a train/test split of 80/20
   For the OneVsRestClassifier and the OneVsOneClassifier use "estimator=GaussianProcessClassifier()".
   Use the proper classes of the scikit learn libs

Literature

https://scikit-learn.org/stable/modules/multiclass.html#ovo-classification
https://scikit-learn.org/stable/modules/multiclass.html
https://scikit-learn.org/stable/modules/generated/sklearn.multiclass.OneVsRestClassifier.html
https://scikit-learn.org/stable/modules/generated/sklearn.multiclass.OneVsOneClassifier.html