Lab Work 3: “Multiclass Classification “

General Remarks

- Please work in groups of two or three people!
- **Clearly mark the code sections that were created with the help of generative AI!**
- Apply train/test data split of **80/20** for every task
- Please use the appropriate sklearn libs

1. Report

- Must be **submitted 3 days before** the presentation
- A written project report (pdf) must contain the results of all tasks (task 1, task 2, task 3 and task 4.1) clearly readable
- Submit the code (pdf), commented and easy to read

2. Presentation

- Show and execute the submitted code
- Explain the tasks and show their results
- **ppt presentation (15 min.) about the results of your findings**

Introduction

The overall goal is to study and assess the performance of various classifiers for a multiclass classification problem on a small-scale data set. For the assessment you will be using the accuracy, precision and the receiver operating Characteristic (ROC) curve. The work consists of two tasks.

The first task deals with the employment of the known multi class solutions techniques “one-vs-rest” or “one-vs-all”[1]. Within this task you also study possible performance enhancements by using Synthetic Minority Over-sampling Technique (SMOTE).

The second sub task deals with the employment of the k-NN classifier[2] [3]. Within this task you study how the k parameter influences the performance of the classifier. Also, you will study the employment SMOTE - Synthetic Minority Over-sampling Technique on the result.

[1] jäger-supervised-decission-reduced-final-2025.pdf; Lecture notes

[2] KNeighborsClassifier — scikit-learn 1.6.1 documentation

[3] jaeger-recomender-final-student-2025.pdf; Lecture notes

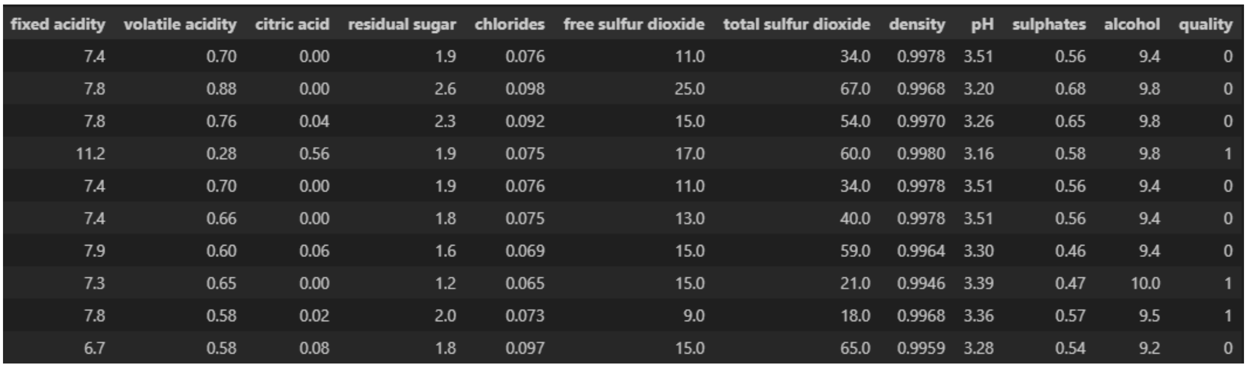
After completion you have learned to

- Apply the one-vs-rest or one-vs-all Classifier for multiclass classification using scikit learn libs
- Apply the k-NN classifier for multi class classification using scikit learn libs
- Assess the obtained accuracy.
- Interpret the ROC curves.
- Work with the SMOTE technique

Data set

You will use the so-called "Wine Quality Dataset" from Kaggle[4]. It is a publicly available data set that contains 1400 samples of a Portuguese red wine with 11 attributes, i.e. physio-chemical data and a quality ranking. The quality is based on sensory data, i.e. not evaluated by humans, and has a range of scores from 1 to 8.

The target Y of the classification task is the quality. To use the quality for a multi class classification task, the original data set the quality was grouped into three groups. The first group contains the "bad" wines, all samples with a quality <=4. The second group contains the "medium" wines with a quality of 5 or 6. And "good" wine with a quality larger than 6. Since in most cases it is easier to work with numbers than with strings, the first group is from now on referred to as quality = 0, the second group as quality = 1 and the third group the good wine as quality = 3. The modified data is stored in the file "**Wine_Test_02_stud.csv**". Fig. 1 shows the first 10 rows of the data set with its 11 attributes and in the last column the changed quality, the target Y, contains for this excerpt only 0s, 1s



fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	0
7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	0
7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	0
11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	1
7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	0
7.4	0.66	0.00	1.8	0.075	13.0	40.0	0.9978	3.51	0.56	9.4	0
7.9	0.60	0.06	1.6	0.069	15.0	59.0	0.9964	3.30	0.46	9.4	0
7.3	0.65	0.00	1.2	0.065	15.0	21.0	0.9946	3.39	0.47	10.0	1
7.8	0.58	0.02	2.0	0.073	9.0	18.0	0.9968	3.36	0.57	9.5	1
6.7	0.58	0.08	1.8	0.097	15.0	65.0	0.9959	3.28	0.54	9.2	0

Figure 1: Excerpt of the modified data set with the attribute "quality".

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

Task 1 One-vs-rest classifier

Tasks 1.1 without SMOTE

a) Display the histograms of all attributes including Y

b) Show the histograms with respect to Y ("0", "1", "2") for all attributes. Please comment on the expected performance of the classifier. Please explain how you came to this conclusion.

c) Perform 10 runs of modeling.

- In each run, split the data set randomly into training and test. Select one estimator e.g.:
model = OneVsRestClassifier(estimator=GaussianProcessClassifier())[5]

- Display the classification report for the test data set for each run.

d) Display the average accuracy with its respective standard deviation of the 10 runs on the test set.

e) Plot the ROC curves (all three classes) for **one run only**

Tasks 1.2 with SMOTE

a) Oversample the data set employing SMOTE[6] such that all three classes should have the same number of samples in the training data. Include SMOTE from imblearn.over_sampling[7]. **SMOTE** is supposed to be applied to the **training dataset** exclusively.

b) Display the histograms of all attributes including Y

c) Show the histograms with respect to Y ("0", "1", "2") for all attributes.

d) Repeat Task1.1 c) bis e) with oversampled trainings data

e) Comment on the results

[5] https://scikit learn.org/stable/modules/generated/sklearn.multiclass.OneVsRestClassifier.html

[6] jäger-supervised-decission-final-student--2025.pdf, Lecture notes

[7] https://imbalanced-learn.org/dev/references/generated/imblearn.over_sampling.SMOTE.html

Task 2: k-NN classifier for a multiclass problem

Tasks 2.1 without SMOTE

The KNeighborsClassifier[8] implements learning based on the k nearest neighbors of each query data sample, where k is an integer value specified by the user.

With the value of k one determines the number of other samples to be considered for the decision. To identify the optimal k parameter of the k-NN classifier the following workflow should be used.

a) Train 15 k-NN classifier, each with a different k value

b) Evaluate each classifier on the test data to determine the best k value. Please, report a plot like Fig. 2.

c) For the best k value perform five runs. Use a random split into training and test data and display the obtained accuracy for each run.

d) Display the average accuracy and its standard deviation.

e) Plot the ROC curves for all three classes for the **best run only.**

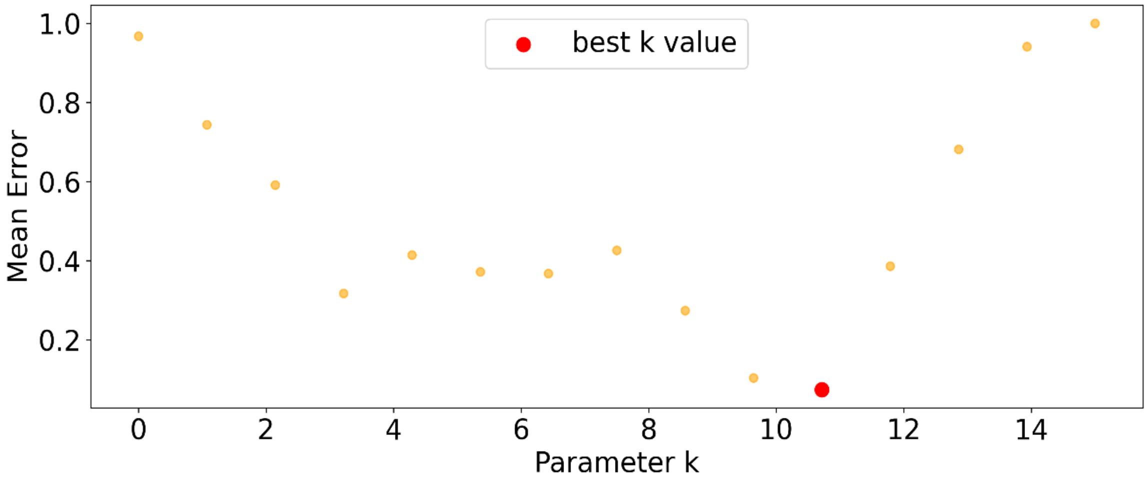


Figure 2: Mean error of the different k-NN classifier with different k parameter evaluated on the test data.

[8] jäger-supervised-decission-final-student--2025.pdf, Lecture notes

Task 2.2 with SMOTE

a) Oversample the data set employing SMOTE

All three classes should have the same number of samples. **SMOTE** is supposed to be applied to the **training dataset**

b) Perform the tasks as stated in Task 2.1 a) bis e)

c) Compare and comment on the results of Task 2.1