**First Exercise**

**Data set description**

Darbą atliko:

Karolis Kvedaravičius

**Vilnius 2025**

# Data set

Chosen data set describes the chemical contents of the Portuguese wine “Vinho Verde” and their subjective (quality) rating from 0 to 10. There is a disbalance between the ratings. There are more “average” wines than highly or poorly rated ones.

There are 1143 different wines in this dataset.

# Attributes

1. fixed acidity
2. volatile acidity
3. citric acid
4. residual sugar
5. chlorides
6. free sulfur dioxide
7. total sulfur dioxide
8. Density
9. pH
10. sulphates
11. alcohol
12. quality (rating between 0 and 10)

The first 11 attributes are nominal attributes that describe the chemical composition of the wine and the 12th attribute is a ordinal attribute that describes the subjective quality of the wine.

The quality distribution of wines:

|  |  |
| --- | --- |
| Rating | Number of wines |
| 3 | 6 |
| 4 | 33 |
| 5 | 483 |
| 6 | 462 |
| 7 | 143 |
| 8 | 16 |

This dataset would be a great candidate for classification task: trying to predict the rating of a wine based on its chemical composition. We can also try to determine which chemical components are the most important to determine the rating.