First Exercise

Data set description

Task performed by:
Karolis Kvedaravičius

Vilnius 2025

Contents

1. 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.

2. Attributes

The data set has 12 attributes and their ranges are listed below.

Table 1 Attributes ant their ranges

	min	max	range
fixed acidity	4.6	15.9	11.3
volatile acidity	0.12	1.58	1.46
citric acid	0	1	1
residual sugar	0.9	15.5	14.6
chlorides	0.012	0.611	0.599
free sulfur dioxide	1	68	67
total sulfur dioxide	6	289	283
density	0.99007	1.00369	0.01362
pH	2.74	4.01	1.27
sulphates	0.33	2	1.67
alcohol	8.4	14.9	6.5
quality	3	8	5

The first 11 attributes are numeric 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. Also, free sulfur dioxide and total sulfur dioxide attributes have much larger numeric values than other attributes. Normalization of the attributes could be needed to perform further statistical analysis.

The quality distribution of wines:

Table 2 wine rating distribution

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 influence the rating the most