Data:
Wine Quality Data Set z UCI Repository of Machine Learning Databases:
http://archive.ics.uci.edu/ml/datasets/Wine+Quality

Popis dat:
Data popisují vlastnosti portugalského vína "Vinho Verde". Jedna sada dat je pro červená a druhá pro bílá vína. Naším úkolem je na základě těchto hodnot stanovit kvalitu vína.

Atributy:

Vstup (acidity = kyselost, citric acid = kyselina citronová)

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

Výstup:

12 - quality (hodnota mezi 0 a 10)

Úkoly: Každému studentovi je přidělena datová sada (bílé nebo červené víno)

Předzpracování dat:

1) Nalezněte průměr, medián, minimum, maximum a standardní odchylku pro tři po sobě jdoucí atributy (například 1-3, nebo 8-10, apod.)
2) Zkonstruujte pro tyto atributy histogramy. Můžete na základě histogramů nalézt nějaké odlehlé hodnoty.
3) Existují mezi těmito atributy korelace?

Strojové učení:

1) Vyberte dvě metody strojového učení a pokuste se klasifikovat kvalitu vína na základě vstupních hodnot.
2) Zkonstruujte křivku učení pro obě vybrané metody a jejich výsledky porovnejte.
3) Zkuste další metodu validace dat.
4) Jaký má vliv nerovnoměrné rozložení tříd na učení?

Shluková analýza:

1) Existují v datech významné podskupiny?
2) Charakterizujte alespoň jednu takovou podskupinu.