Assessing wine quality

COURSERA ADVANCED CAPSTONE

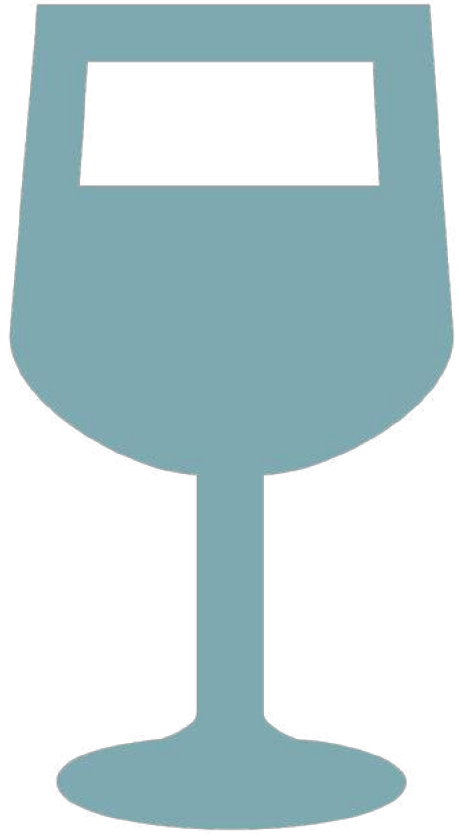
Vineyards always collect and sample their products

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5

Dataset contains 1600 wine batches, chemically analysed (Cortez et al., 2009)

All stakeholders to benefit from automatic quality assessment

- Today, wine batches are quality assessed manually, taking time and making them susceptible to bias
- With our model, this can be made automatic based solely on the physiochemical properties(!)



Architectual desicions

JUPYTER NOTEBOOKS

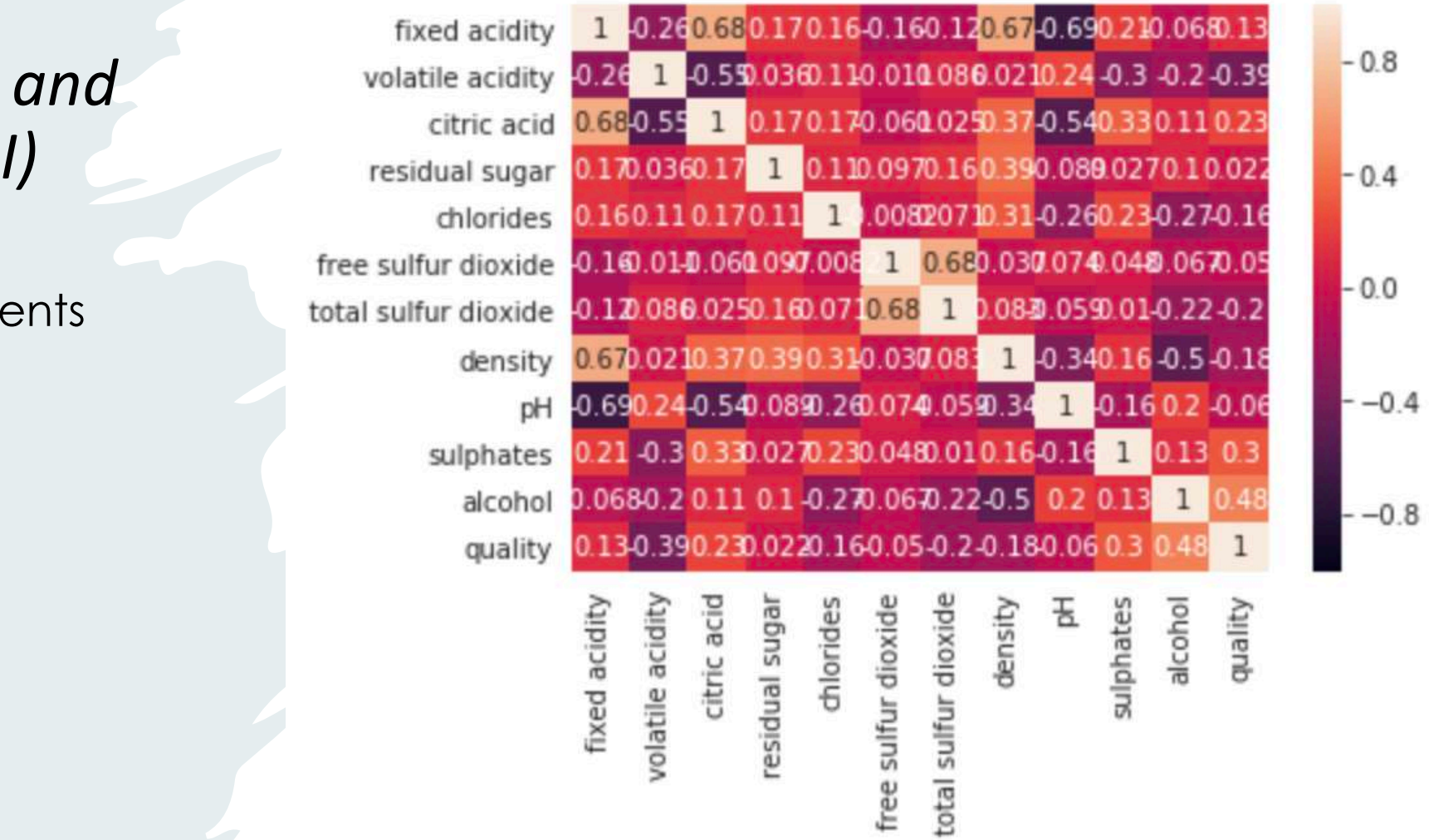
GITHUB REPOSITORY

Deep-dive – Data structure, cleaning and pre-processing (I/II)

- Dataset contain 1599 rows (batches), each with 11 features and 1 target variable
  - All features numerical (floats) – Outliers removed and data normalized (z-score normalization)
  - Target state categorical (0-10) – One-hot encoded to either 0/1 (bad/good) as
    - <6 = 0
    - >=6 = 1

Deep-dive – Data structure, cleaning and pre-processing (II/II)

- Correlation measurements and individual feature feasibility studied



Model selection & performance indicators

- Binary classification problem (0 Bad/1 Good)
  - Logistic regression
  - Support Vector Machines
- Highly tuneable, GridSearch used for parameter sweeps
- Accuracy main performance metric
  - Confusion matrices
  - Correlation measures (F1)

Results – Best performing SVM

Used the 'rbf' kernel with C=10 regularization

Overall accuracy of 90%

Precision 96% (0/Bad) & 71% (1/Good)

Outlook

- Our model confidently predicts bad wines from good ones, but is not as good in the opposite (due to there being much more bad data points than good)
- The models resulting from our (limited) 1600 examples, we are able to match human prediction **>90%** of the time
- Final deliverables include the model along with its documentation