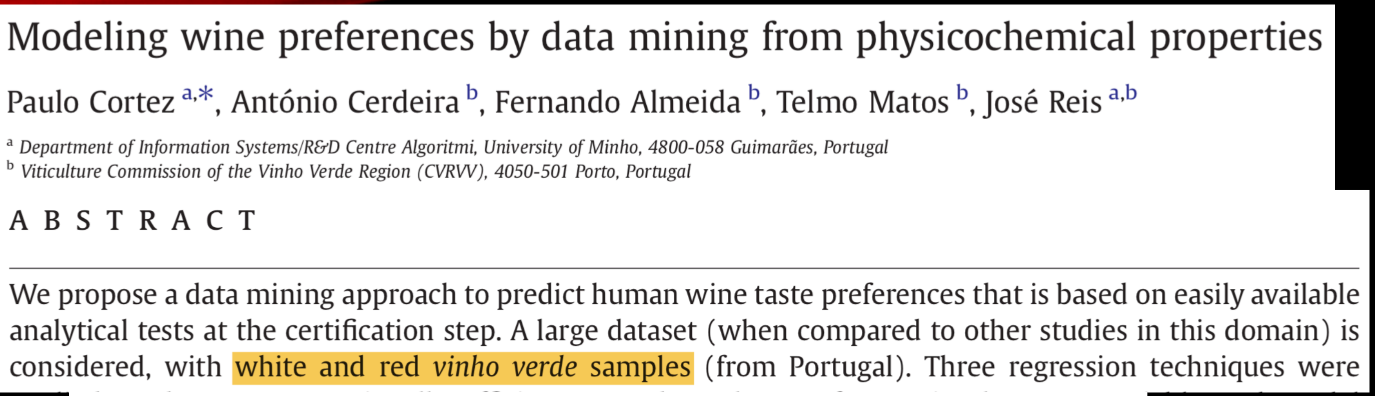
ADVANCED PRACTICAL COURSE DATA SCIENCE TASK 1 – FINAL PRESENTATION

Anonymized to be used on github.com/falo0

17-05-2018

RESEARCH QUESTION AND DATA



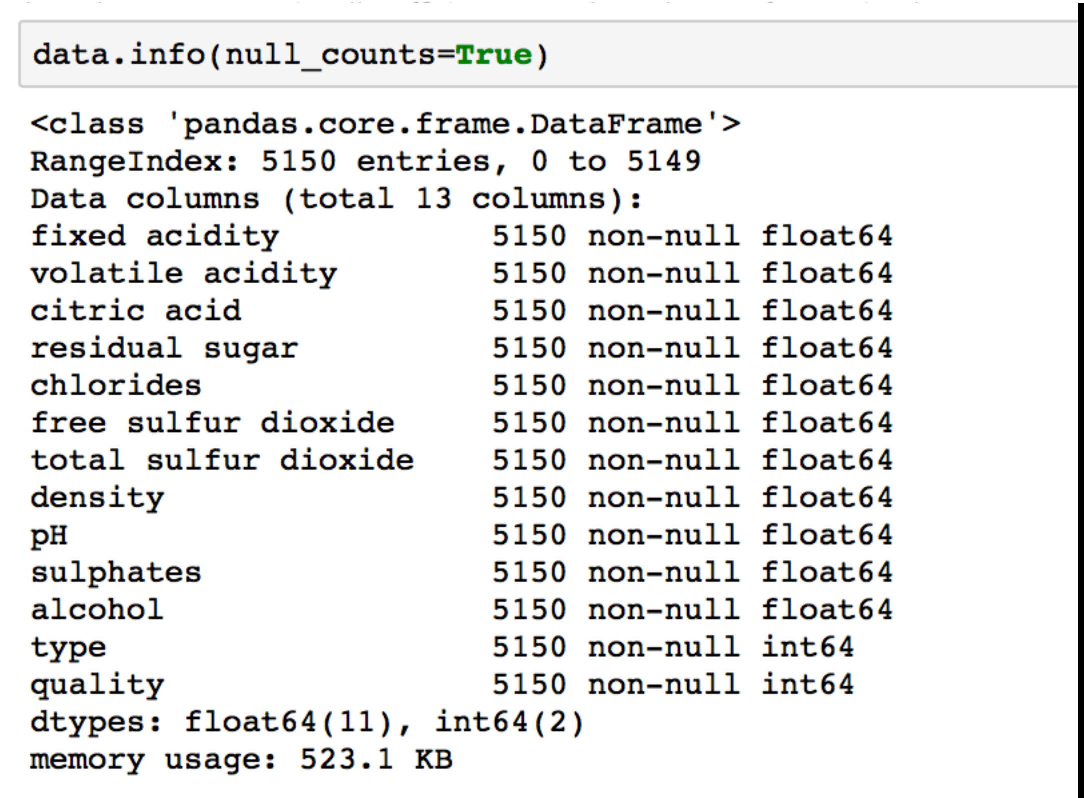
Modeling wine preferences by data mining from physicochemical properties

Paulo Cortez [a,*], António Cerdeira [b], Fernando Almeida [b], Telmo Matos [b], José Reis [a,b]

[a] *Department of Information Systems/R&D Centre Algoritmi, University of Minho, 4800-058 Guimarães, Portugal*
[b] *Viticulture Commission of the Vinho Verde Region (CVRVV), 4050-501 Porto, Portugal*

ABSTRACT

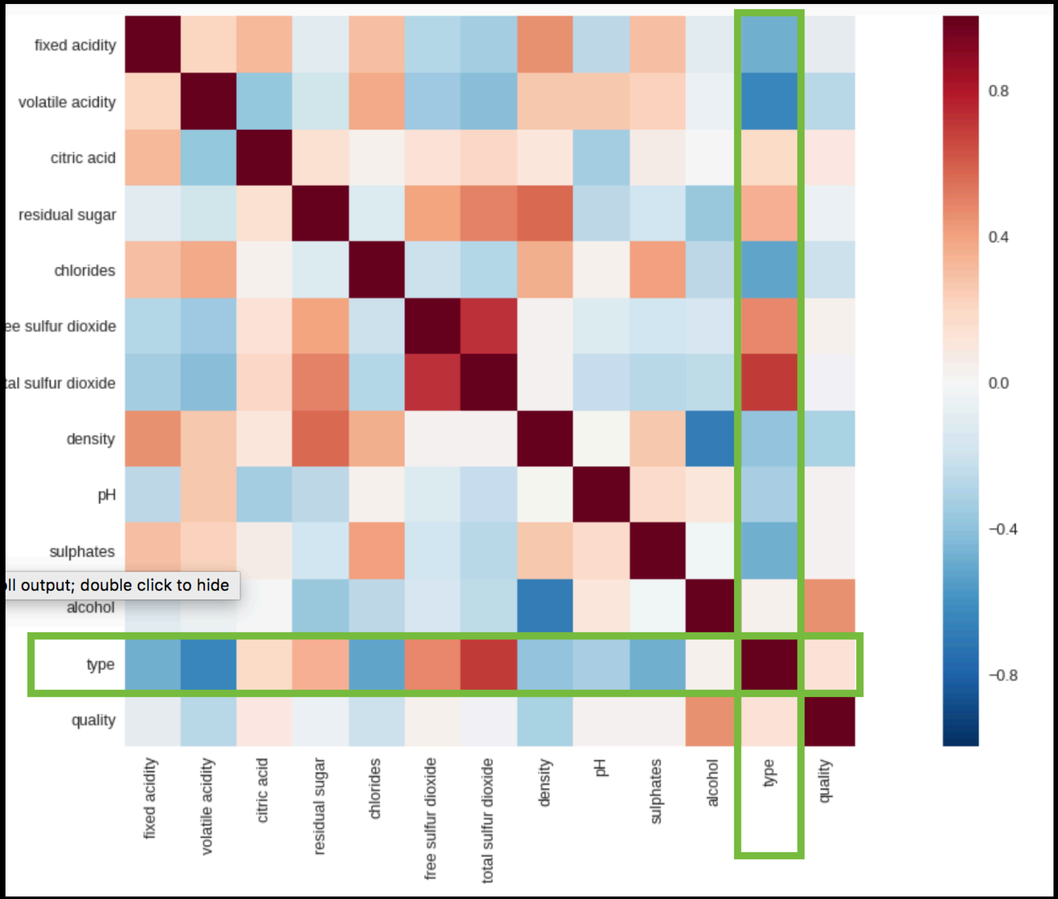
We propose a data mining approach to predict human wine taste preferences that is based on easily available analytical tests at the certification step. A large dataset (when compared to other studies in this domain) is considered, with white and red *vinho verde* samples (from Portugal). Three regression techniques were



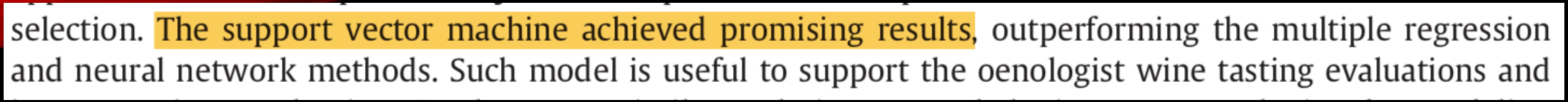
```
data.info(null_counts=True)
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5150 entries, 0 to 5149
Data columns (total 13 columns):
fixed acidity           5150 non-null float64
volatile acidity        5150 non-null float64
citric acid             5150 non-null float64
residual sugar          5150 non-null float64
chlorides               5150 non-null float64
free sulfur dioxide     5150 non-null float64
total sulfur dioxide    5150 non-null float64
density                 5150 non-null float64
pH                      5150 non-null float64
sulphates               5150 non-null float64
alcohol                 5150 non-null float64
type                    5150 non-null int64
quality                 5150 non-null int64
dtypes: float64(11), int64(2)
memory usage: 523.1 KB
```

Feature Correlations



DATA SCIENCE TOOLS



selection. The support vector machine achieved promising results, outperforming the multiple regression and neural network methods. Such model is useful to support the oenologist wine tasting evaluations and

<u>Tried so far:</u>

sklearn.svm with default settings

- Kaggle score of 0.49009

sklearn.ensemble.RandomForestRegressor with 50 trees for red and 200 trees for white

- Kaggle score of 0.47524

<u>To Do:</u>

- More thought about feature selection
- Try different settings for the SVM and the RF
- Maybe try another Data Science tool, like a neural network, even though they coudn’t achieve best results with a NN in the paper
- Maybe construct a meta model of the different tools/models used