

# **White Wine Quality**

group WOE Ran Huang, Zitong Zeng, Tian Qi, Randy Ma

#### **Data Overview**



- The data is Vinho Verde\* white wine samples from Portugal.
- The goal of this project is to determine wine quality based on the chemical features (Cortez et al., 2009)
  - Input variable: based on physicochemical tests
  - Output variable: based on sensory data, median of at least 3 evaluations made by wine experts



#### **Features**

- Fixed acidity
- Volatile acidity
- Citric acid
- Residual sugar
- Chlorides
- Free sulfur dioxide
- Total sulfur dioxide
- Density
- PH
- Sulphates
- Alcohol

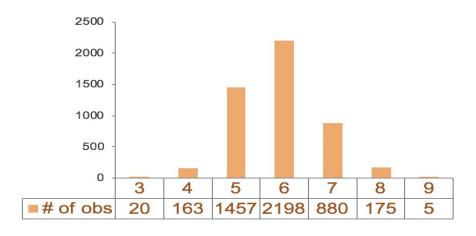


## Labels and Encoding

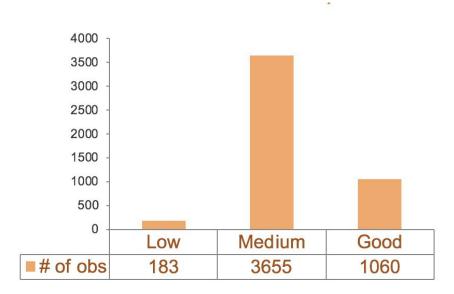
- Quality is represented by scores ranging from 0 to 10
- 0 is the worst and 10 is the best
- Relabel:
  - score under 5 → "Low"
  - score above 6 → "High"
  - score of 5 and 6 → "Medium"



## **Data Distribution and Encoding**



- Score under 5 → "Low"
- score above 6 → "High"
- o score of 5 and 6 → "Medium"



# **Correlation Matrix of Features**



fixed acidity -	1	-0.023	0.29	0.089	0.023	-0.049	0.091	0.27	-0.43	-0.017	-0.12	-0.11
volatile acidity -	-0.023	1	-0.15	0.064	0.071	-0.097	0.089	0.027	-0.032	-0.036	0.068	-0.19
citric acid	0.29	-0.15	1	0.094	0.11	0.094	0.12	0.15	-0.16	0.062	-0.076	-0.0092
residual sugar -	0.089	0.064	0.094	1	0.089	0.3	0.4	0.84	-0.19	-0.027	-0.45	-0.098
chlorides -	0.023	0.071	0.11	0.089	1	0.1	0.2	0.26	-0.09	0.017	-0.36	-0.21
free sulfur dioxide -	-0.049	-0.097	0.094	0.3	0.1	1	0.62	0.29	-0.00062	0.059	-0.25	0.0082
total sulfur dioxide -	0.091	0.089	0.12	0.4	0.2	0.62	1	0.53	0.0023	0.13	-0.45	-0.17
density -	0.27	0.027	0.15	0.84	0.26	0.29	0.53	1	-0.094	0.074	-0.78	-0.31
pH -	-0.43	-0.032	-0.16	-0.19	-0.09	-0.00062	0.0023	-0.094	1	0.16	0.12	0.099
sulphates -	-0.017	-0.036	0.062	-0.027	0.017	0.059	0.13	0.074	0.16	1	-0.017	0.054
alcohol -	-0.12	0.068	-0.076	-0.45	-0.36	-0.25	-0.45	-0.78	0.12	-0.017	1	0.44
quality -	-0.11	-0.19	-0.0092	-0.098	-0.21	0.0082	-0.17	-0.31	0.099	0.054	0.44	1
	fixed acidity -	volatile acidity -	citric acid -	residual sugar -	chlorides -	free sulfur dioxide –	total sulfur dioxide –	density -	౼	sulphates -	alcohol -	quality -

-0.9

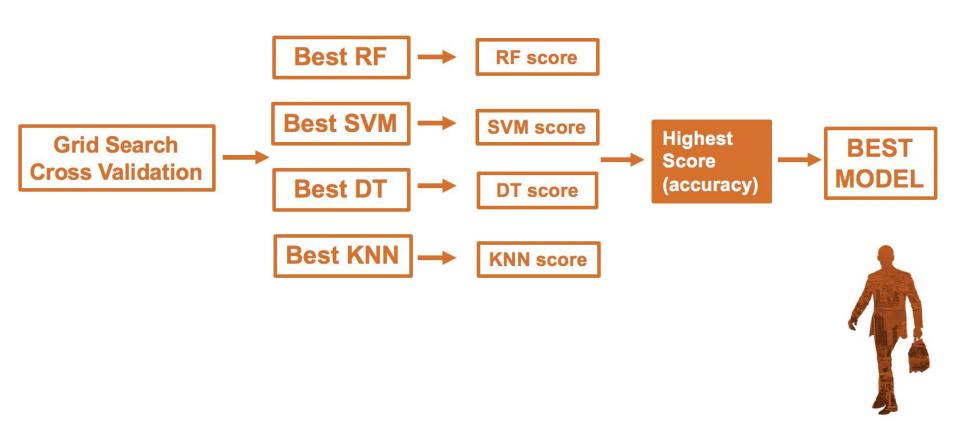
- 0.6

- 0.3

- 0.0

# **Modeling Process**





## **Validation Accuracy**



#### accuracy

Mean cross-validated score (accuracy) of the best estimator



Random Forest performs the best on validation

## **RF on Testing Accuracy**



		precision	recall	f1-score	support
	0	0.79	0.61	0.69	209
	1	0.57	0.11	0.19	35
	2	0.86	0.95	0.90	736
micro	avg	0.85	0.85	0.85	980
macro	avg	0.74	0.56	0.59	980
weighted	avg	0.84	0.85	0.83	980

The RF model accuracy on Test data is 0.8469387755102041

## **RF Confusion Matrix**



## **Confusion Matrix**

#### Of Random Forest

	Predicted	GOOD	LOW	MEDIUM
Actual	GOOD	128	0	81
	LOW	0	4	31
	MEDIUM	35	3	698



## Resample

The data is significantly imbalanced, so we decided to resample to improve accuracy on low and high

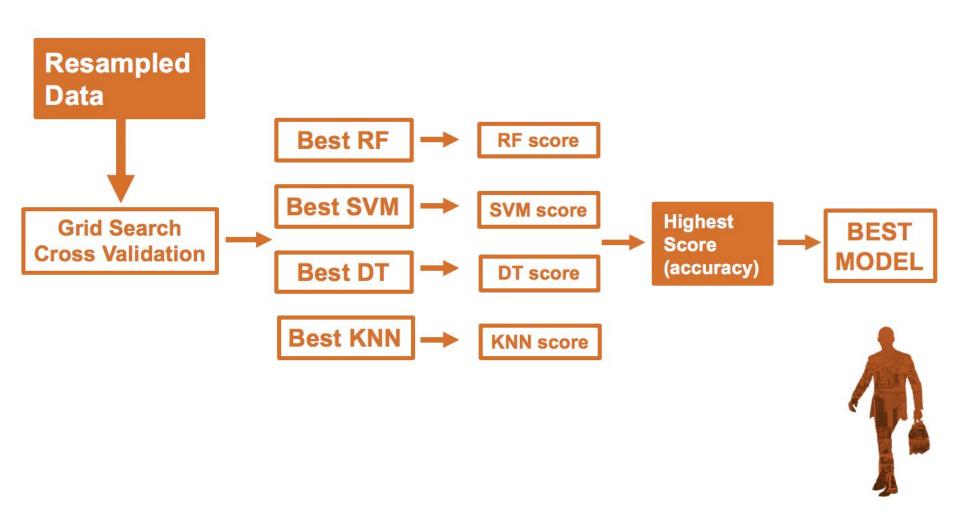
LOW 148 LOW 1500

MEDIUM 2919 MEDIUM 1500

GOOD 851 GOOD 1500

#### **Similar Process**





## **Validation Accuracy**



#### accuracy

Mean cross-validated score (accuracy) of the best estimator (using resampled data)



Random Forest is still the best. SVM also performs well with resampled data.



## **RF Confusion Matrix**

RF acc on test data BEFORE resampling: 84.69%

	Predicted	GOOD	LOW	MEDIUM
Actual	GOOD	128	0	81
	LOW	0	4	31
	MEDIUM	35	3	698

RF acc on test data AFTER resampling: 73.87%

5	Predicted	GOOD	LOW	MEDIUM
Actual	GOOD	174	0	35
	LOW	2	11	22
	MEDIUM	151	46	539



## RF vs. SVM

RF acc on test data AFTER resampling: 73.87%

	Predicted	GOOD	LOW	MEDIUM
Actual	GOOD	174	0	35
	LOW	2	11	22
	MEDIUM	151	46	539

**SVM** acc on test data **AFTER** resampling: 78.88%

	Predicted	GOOD	LOW	MEDIUM
Actual	GOOD	106	4	99
	LOW	0	7	28
	MEDIUM	49	27	660

#### Reference



http://www3.dsi.uminho.pt/pcortez/wine/

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis.

Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.