
CHATEAU DE
LEELANAU

White Wine Quality

Presented by

Wildest On Earth

Agenda

-- Data Analysis & Feature Engineering

-- Modeling & Metric Evaluation

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 properties**

  - Input variable: based on physicochemical tests

  - Output variable: based on sensory data, median of at least 3 evaluations made by wine experts

*Portuguese wine that originated in the historic Minho province in the far north of the country

Features

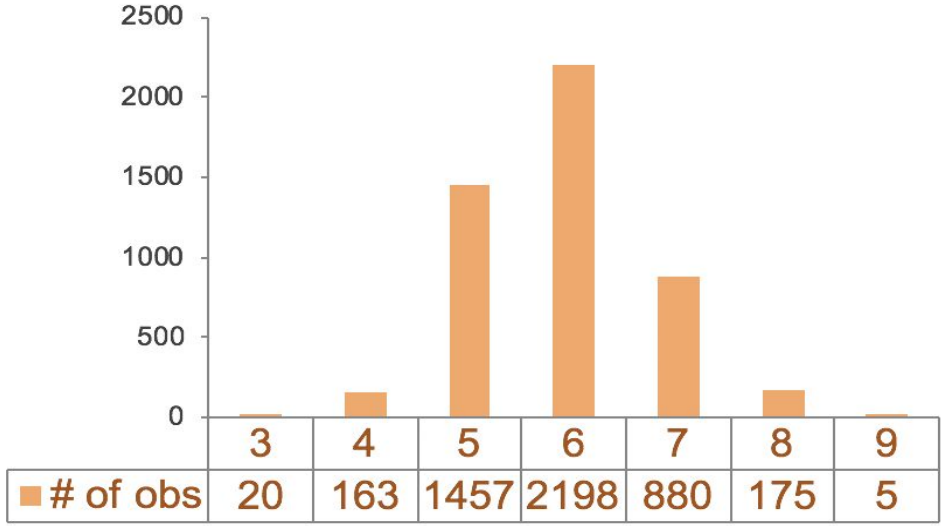
- Fixed acidity
- Volatile acidity
- Citric acid
- Residual sugar
- Chlorides
- Free sulfur dioxide
- Total sulfur dioxide
- Density
- PH
- Sulphates
- Alcohol

```
wine.isnull().sum()
```

```
fixed acidity           0
volatile acidity        0
citric acid             0
residual sugar          0
chlorides               0
free sulfur dioxide     0
total sulfur dioxide    0
density                 0
pH                      0
sulphates               0
alcohol                 0
quality                 0
```

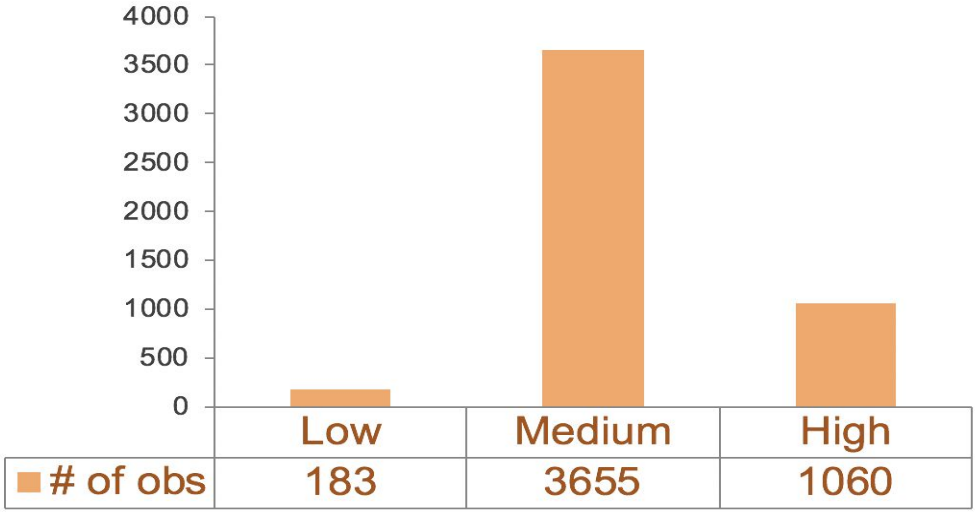
Label Distribution

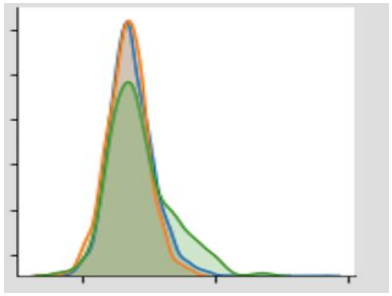
- Quality is represented by scores ranging from 0 to 10
- 0 is the worst and 10 is the best



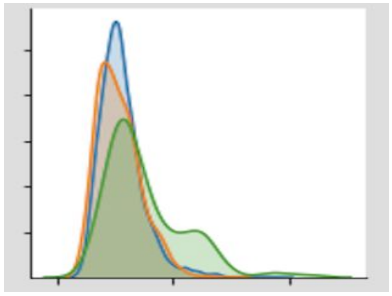
Labels and Encoding

- Binning:
  - score under 5 → “Low”
  - score above 6 → “High”
  - score of 5 and 6 → “Medium”

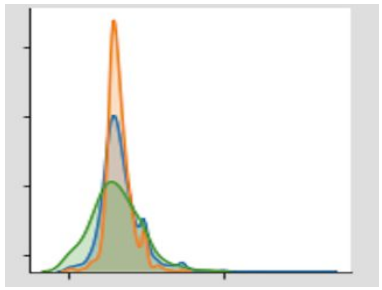




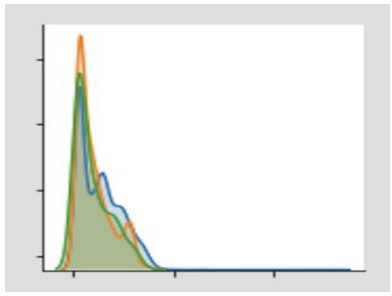
fixed acidity



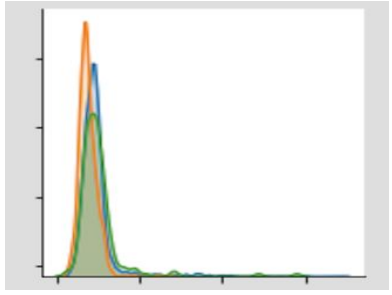
volatile acidity



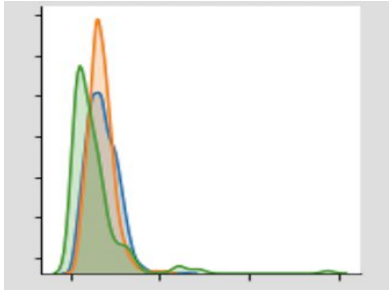
citric acid



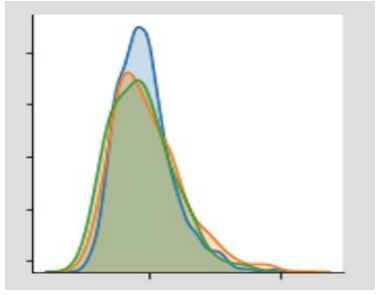
residual sugar



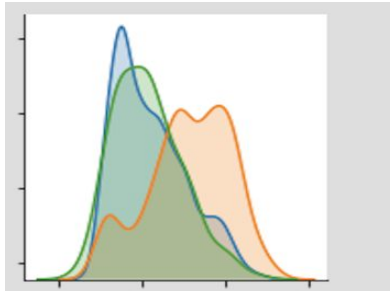
chlorides



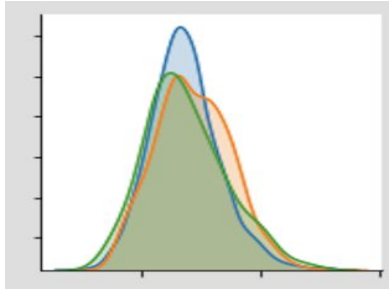
free sulfur dioxide



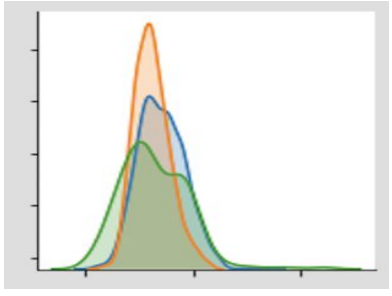
sulphates



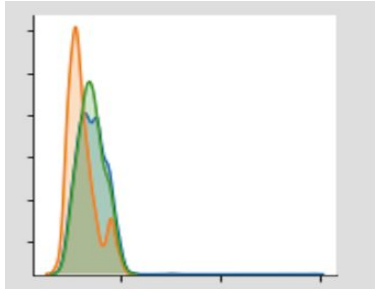
alcohol



PH

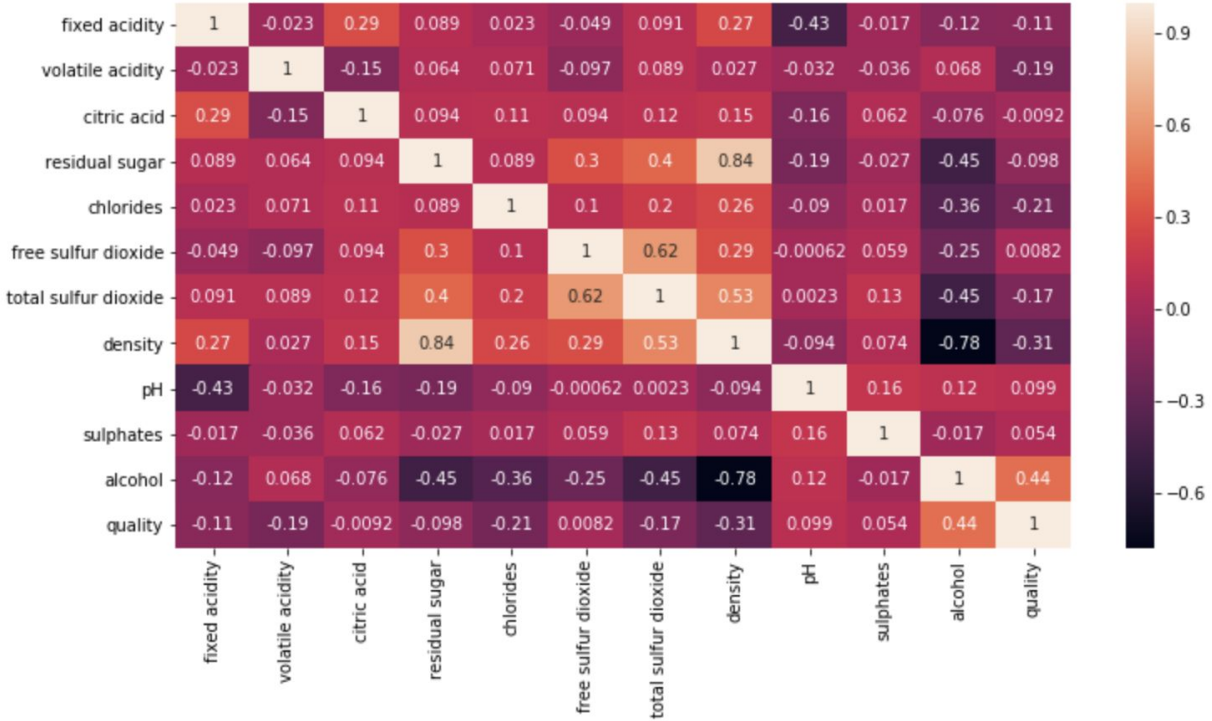


total sulfur dioxide

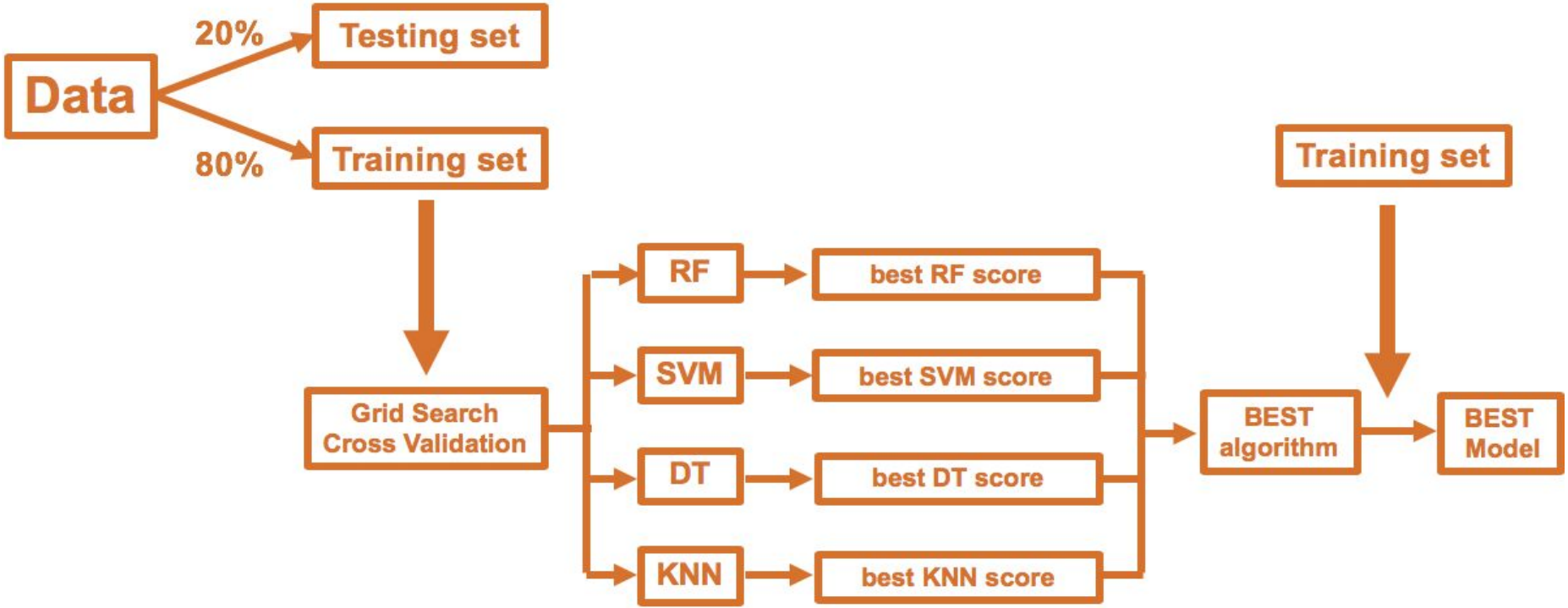


density

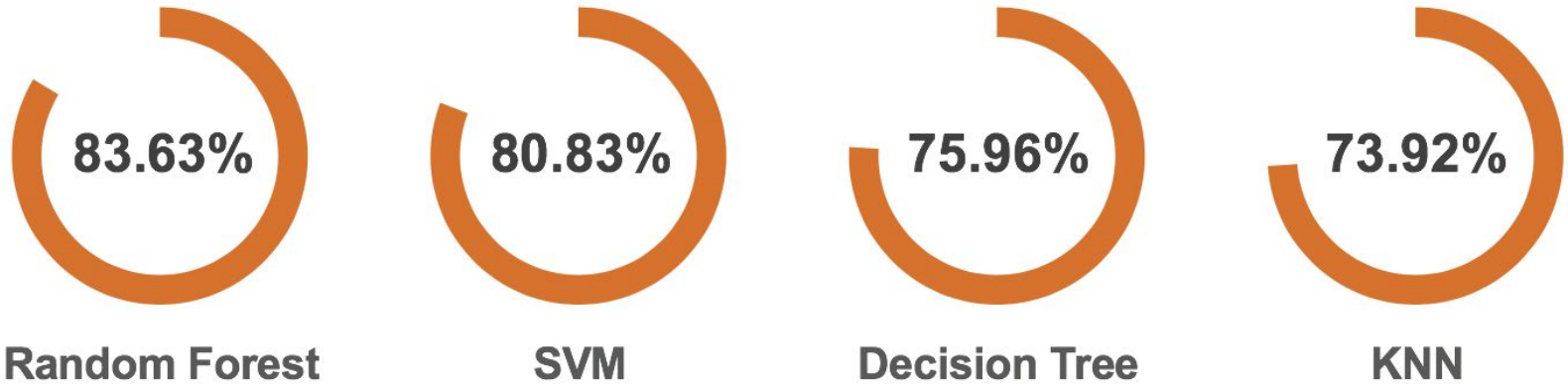
Variable Correlation Matrix



Modeling Process



Validation Accuracy



Random Forest performs the best on validation

RF on Testing Accuracy

Quality	Precision	Recall	F1-score	Actual Count
High	0.79	0.61	0.69	209
Low	0.57	0.11	0.19	35
Medium	0.86	0.95	0.90	736
Weighted Average	0.84	0.86	0.83	980
Model Accuracy on Test Data			**0.85**	

RF Confusion Matrix

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

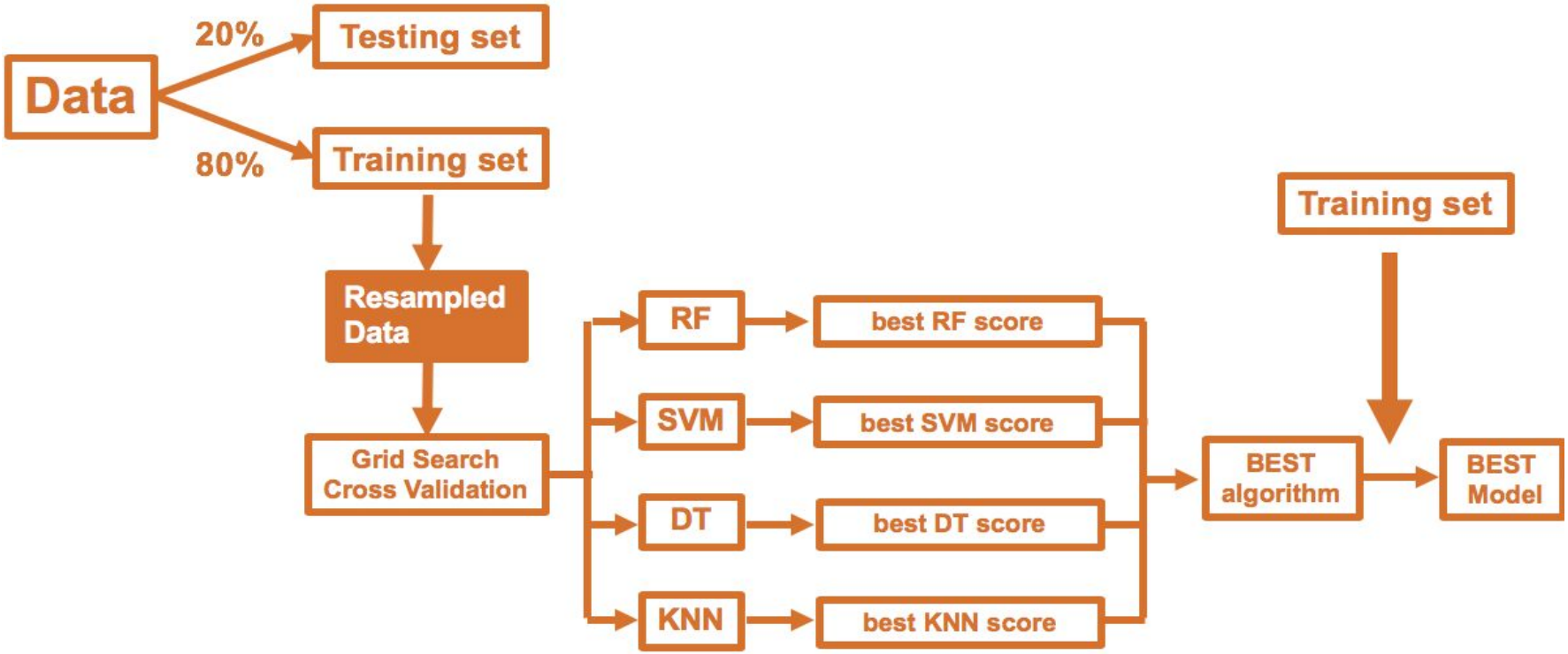
Resample

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

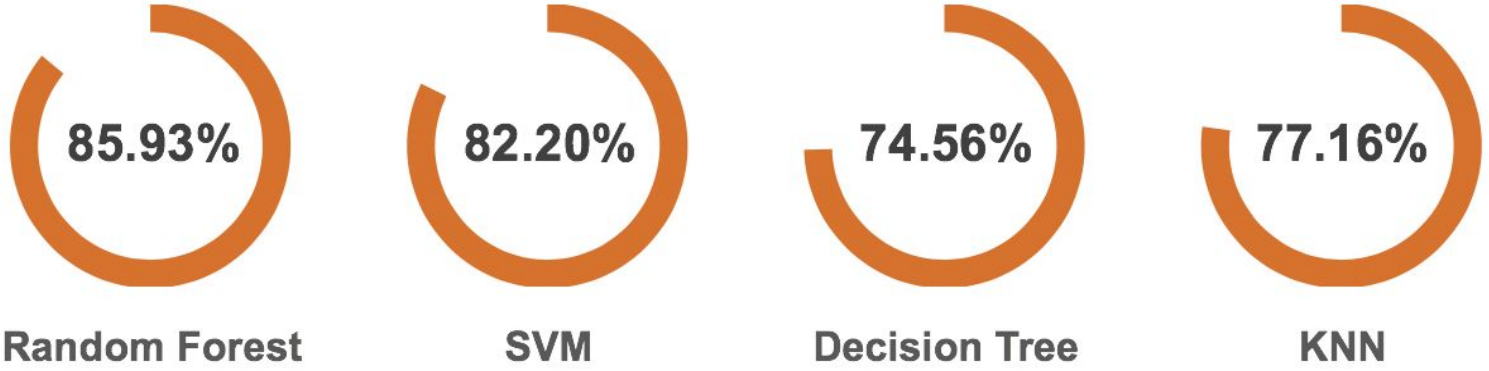
LOW	148
MEDIUM	2919
HIGH	851

LOW	1500
MEDIUM	1500
HIGH	1500

Similar Process



Validation Accuracy



Random Forest is still the best.

RF Confusion Matrix

RF acc on test data BEFORE resampling: 84.69%

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

RF acc on test data AFTER resampling: 73.87%

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

RF vs. SVM

RF acc on test data AFTER resampling: 73.87%

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

SVM acc on test data AFTER resampling: 78.88%

	Predicted	HIGH	LOW	MEDIUM
Actual	HIGH	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.

Thank you ！

Q&A

SVM Confusion Matrix

SVM acc on test data BEFORE resampling: 82.65%

	Predicted	HIGH	LOW	MEDIUM
Actual	HIGH	72	0	137
	LOW	0	2	33
	MEDIUM	0	0	736

SVM acc on test data AFTER resampling: 78.88%

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