# 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

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

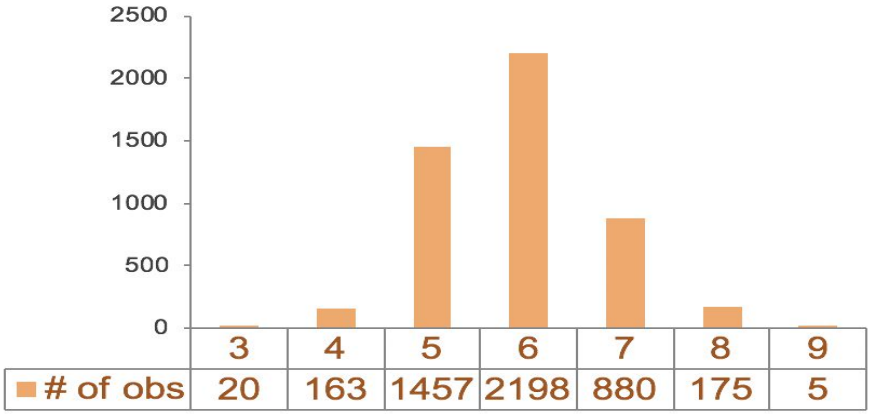
# Features

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

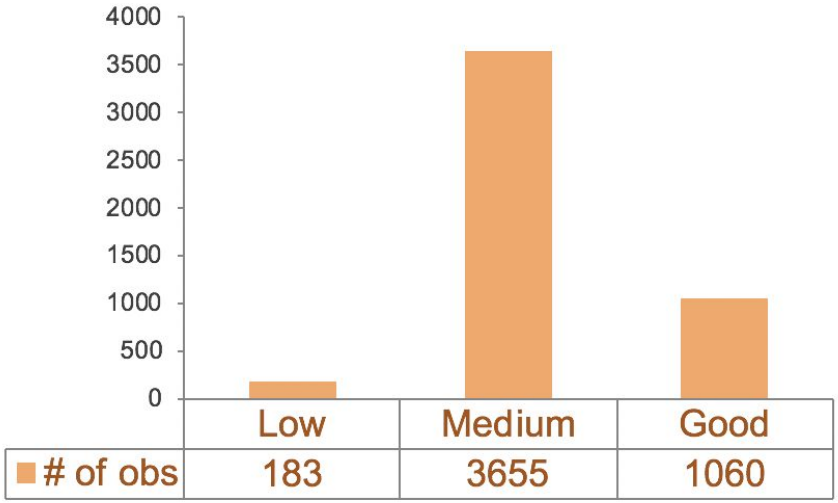
# Labels and Encoding

- 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”

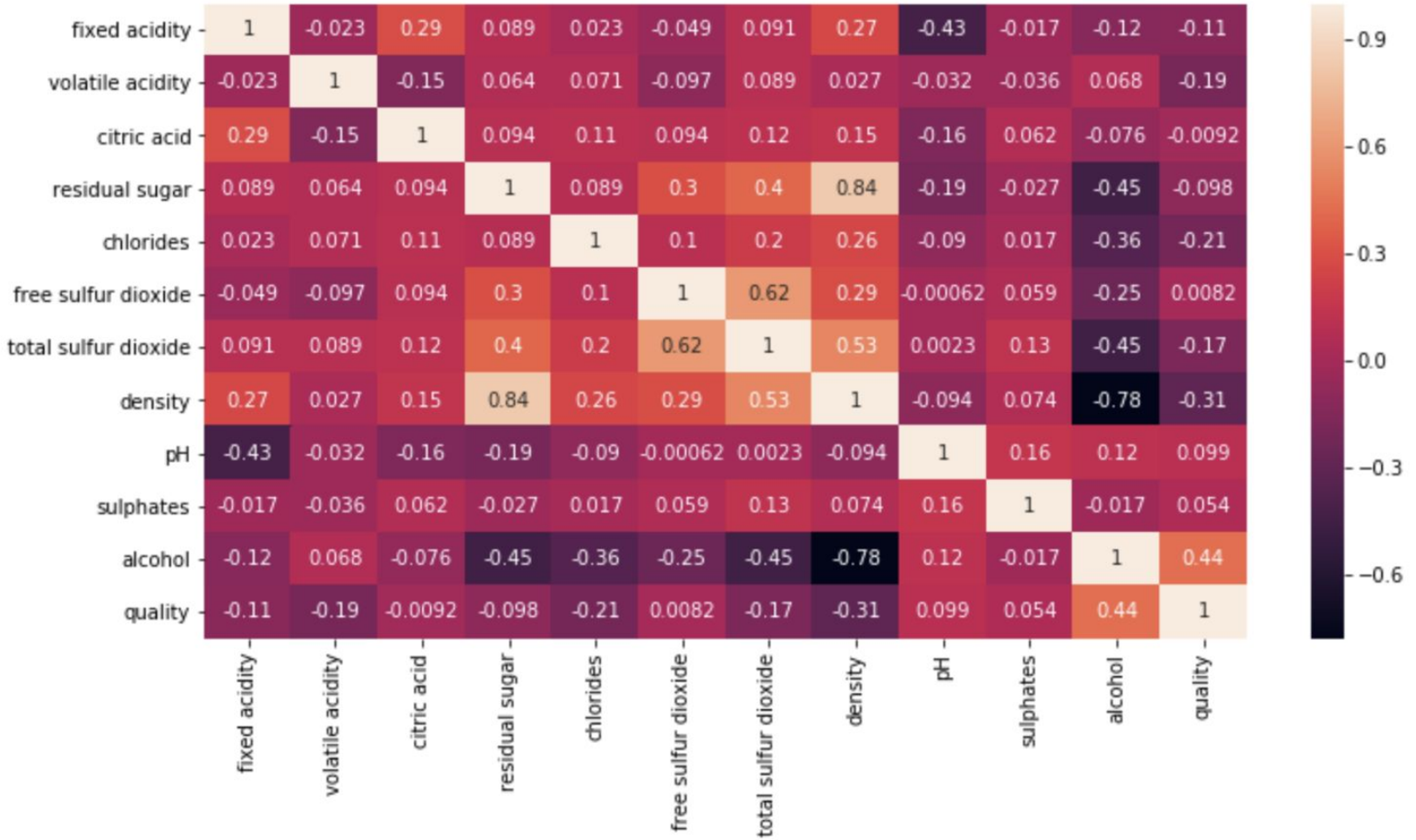
# Data Distribution and Encoding




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



# Correlation Matrix of Features

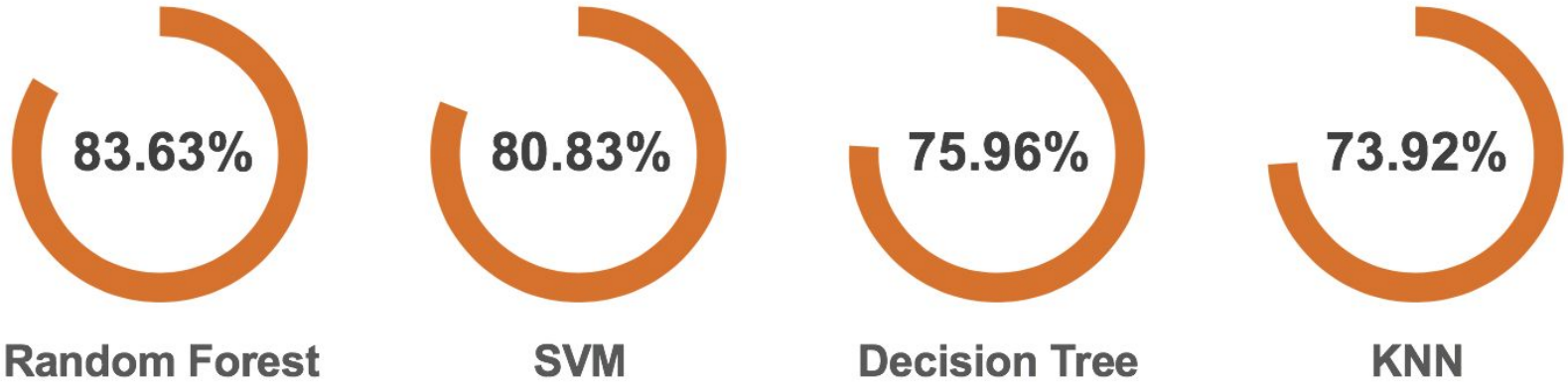


# Modeling Process

# 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

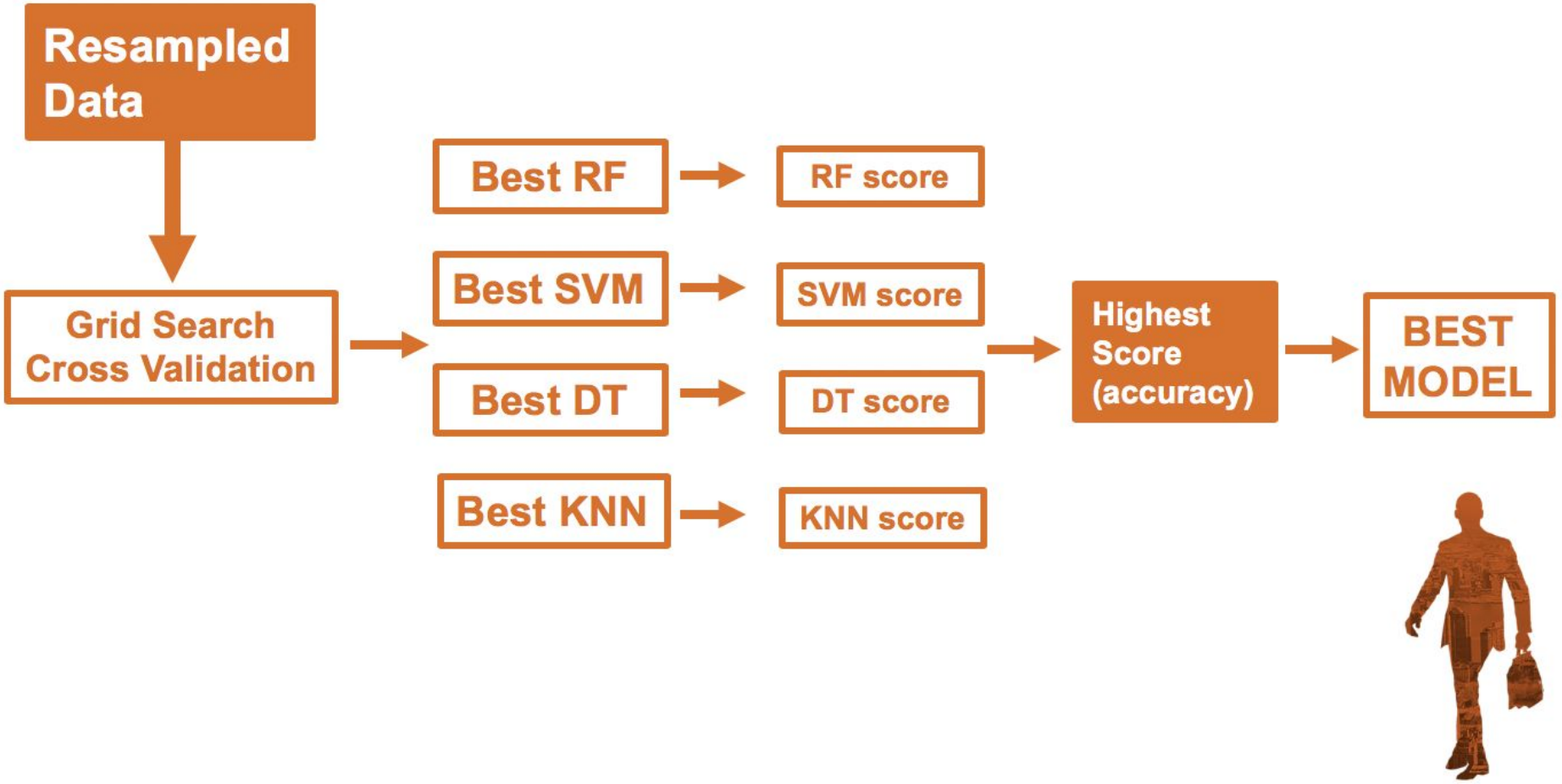
| | Predicted | GOOD | LOW | MEDIUM |
|---|---|---|---|---|
| **Actual** | GOOD | 128 | 0 | 81 |
| | LOW | 0 | 4 | 31 |
| | MEDIUM | 35 | 3 | 698 |

# 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 |

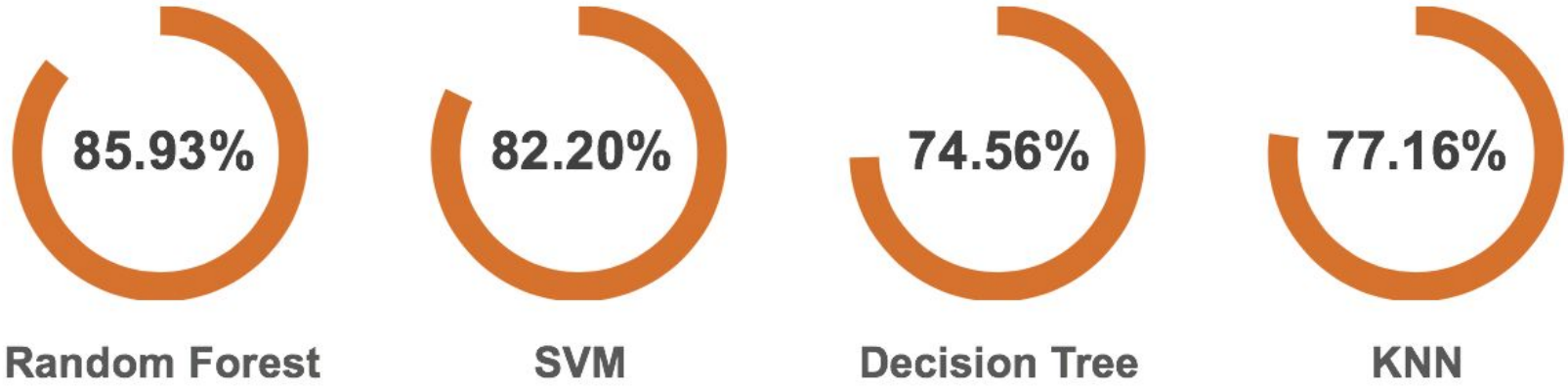
# Similar Process



# 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%**

| | 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.