# Predicting *Vinho Verde* Quality using Random Forests

*Megan McGoldrick*

*GA Data Science*

*April 2015*

# Objective

In 2009, a research team in Portugal showed that SVM outperformed NN and MLR methods in “predict[ing] human wine taste preferences based on easily available analytical tests.” [a]

My goal is to achieve ~~comparable or~~ **better** classification performance with less effort on data transformation, parameter tuning and/or feature selection.

[a] 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.

# Data Source

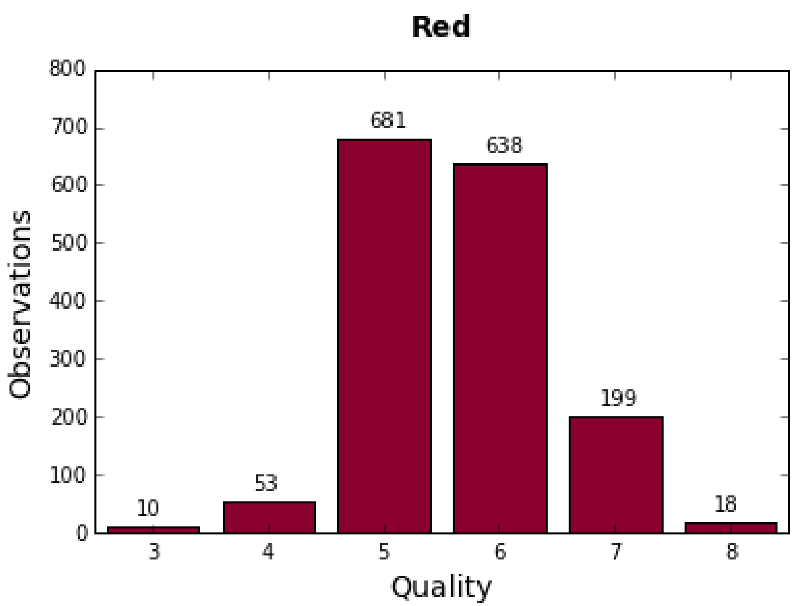
- Samples of *Vinho Verde* (“young wine”) from northwest Portugal
  – 1,599 red, 4,898 white
- Features
  – 1 response: sensory-based measure of quality
  – 11 explanatory (physiochemical properties): Fixed acidity, volatile acidity, citric acid, pH, alcohol, residual sugar, chlorides, density, sulphates, free sulfur dioxide and total sulfur dioxide
- No missing values
- Collected by Portugal’s official certification body (CVRVV) between May 2004 and February 2007
- Available in UCI MLR
  [http://archive.ics.uci.edu/ml/datasets/Wine+Quality](http://archive.ics.uci.edu/ml/datasets/Wine+Quality)

# Multiple, imbalanced classes

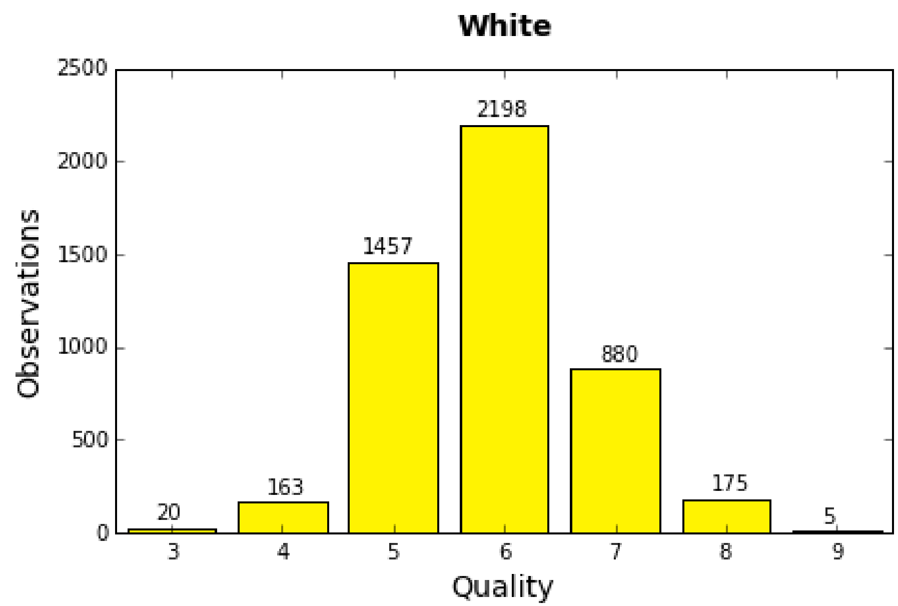
Most samples perceived as “average;” few high/low quality

## Distribution of Quality

*Median of 3+ expert ratings on 0-10 scale from blind tastings*

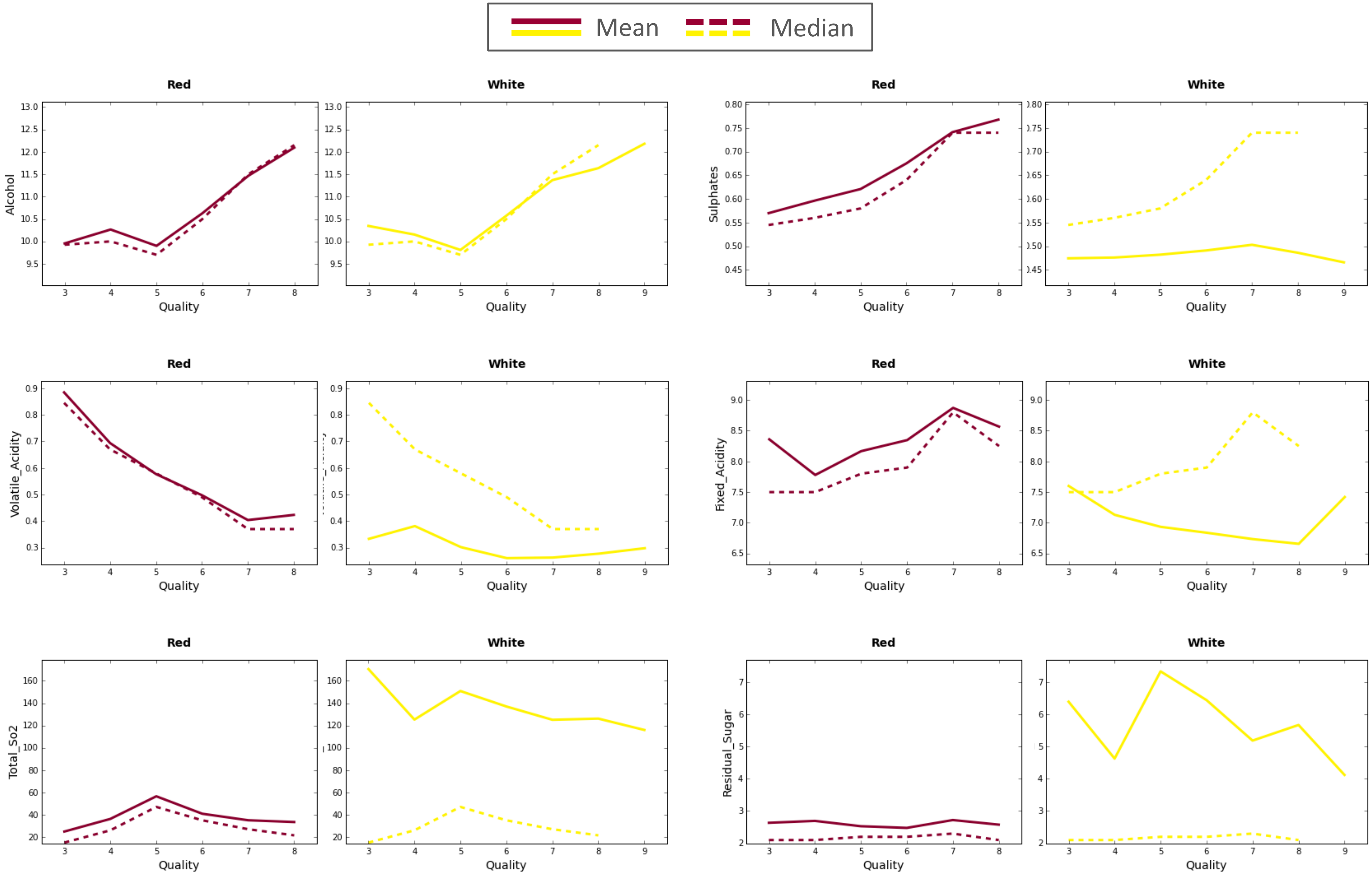



**Mean** 5.6 **Median** 6

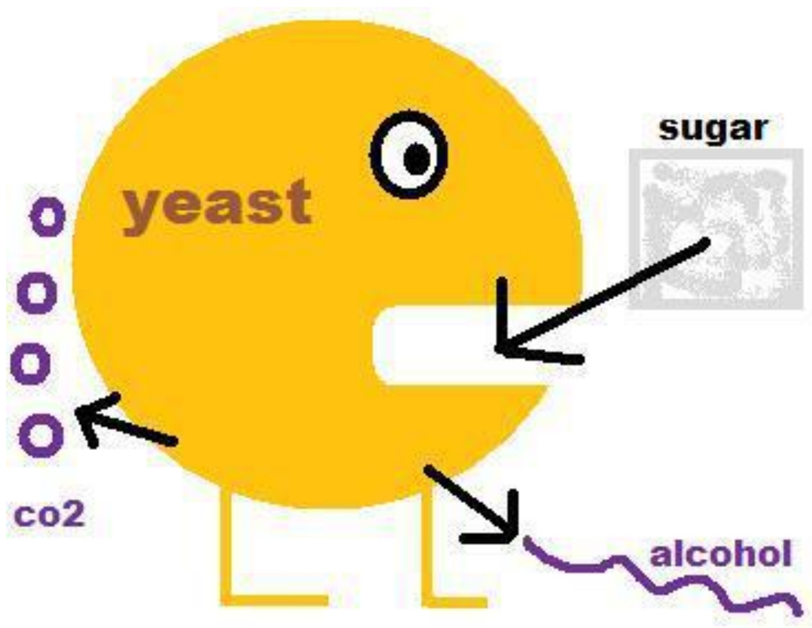



**Mean** 5.9 **Median** 6

# Skewed/noisy data, non-linear relationships



# Dependent/correlated features




**Selected Spearman Rank Correlation Coefficients**
(p-value < .0001)

| | Red | White |
|---|---|---|
| Density - Alcohol | -0.5 | -0.8 |
| Density - Residual Sugar | 0.4 | 0.8 |
| Density - Chlorides | 0.4 | 0.5 |
| pH - Fixed Acidity | -0.7 | -0.4 |
| pH - Citric Acid | -0.5 | |
| Free S02 - Total S02 | 0.8 | 0.6 |
| Citric Acid - Fixed Acidity | 0.7 | |
| Citric Acid - Volatile Acidity | -0.6 | |

# Modeling Approach

## Separate for red and white

1. 20 iterations of Scikit-Learn's Random Forest Classifier with 500 trees and stratified 5-fold cross validation

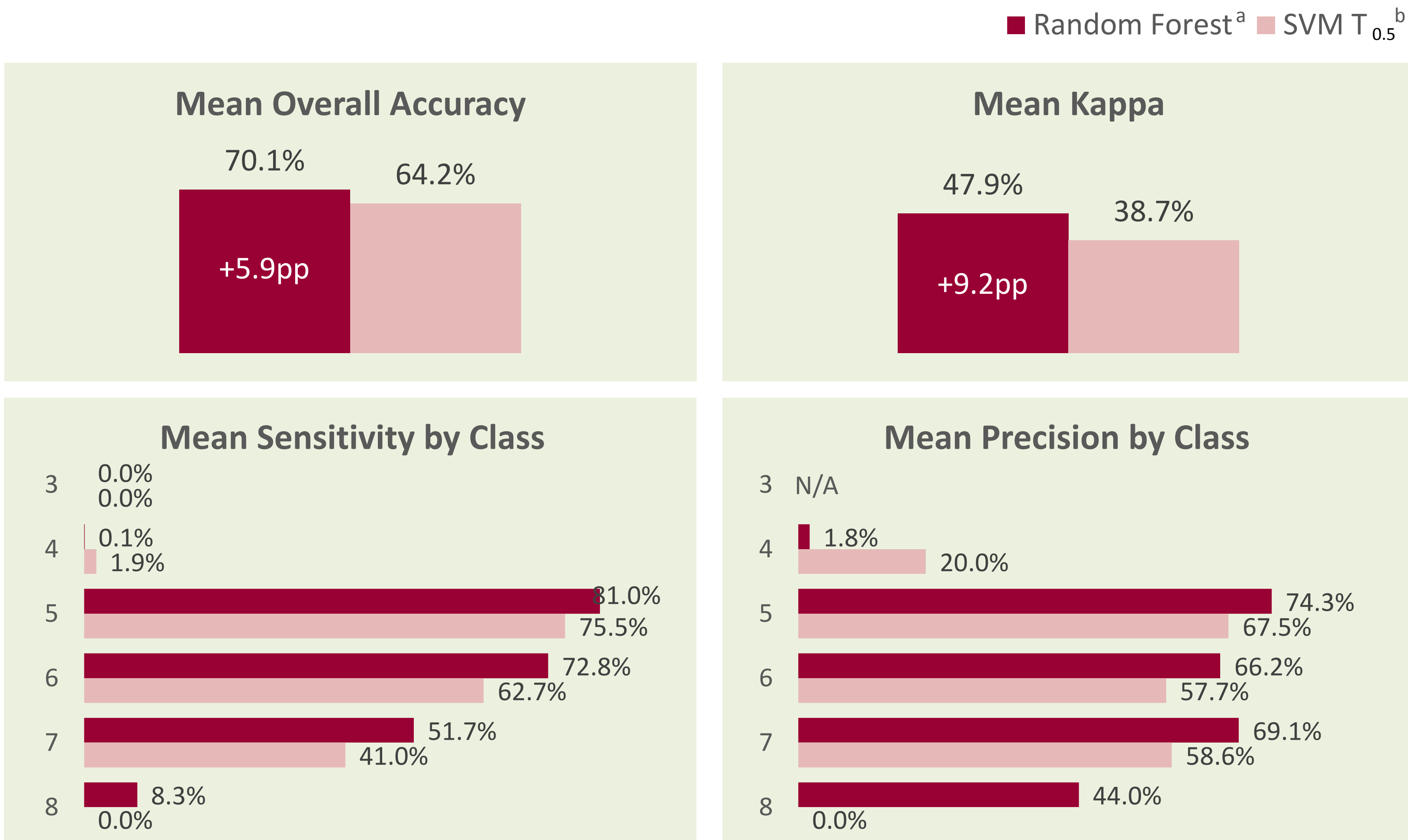
2. Grid search with F1 scoring to tune parameters for number of trees, max features per tree and measure of tree split quality (gini vs entropy)

3. 20 additional iterations of Random Forest Classifier with "best" model parameters and stratified 5-fold CV

4. Comparison of RF overall accuracy, kappa, sensitivity by class and precision by class to published SVM $T_{0.5}$ results

# Red: RF shows some gains over SVM

## Particularly for higher classes

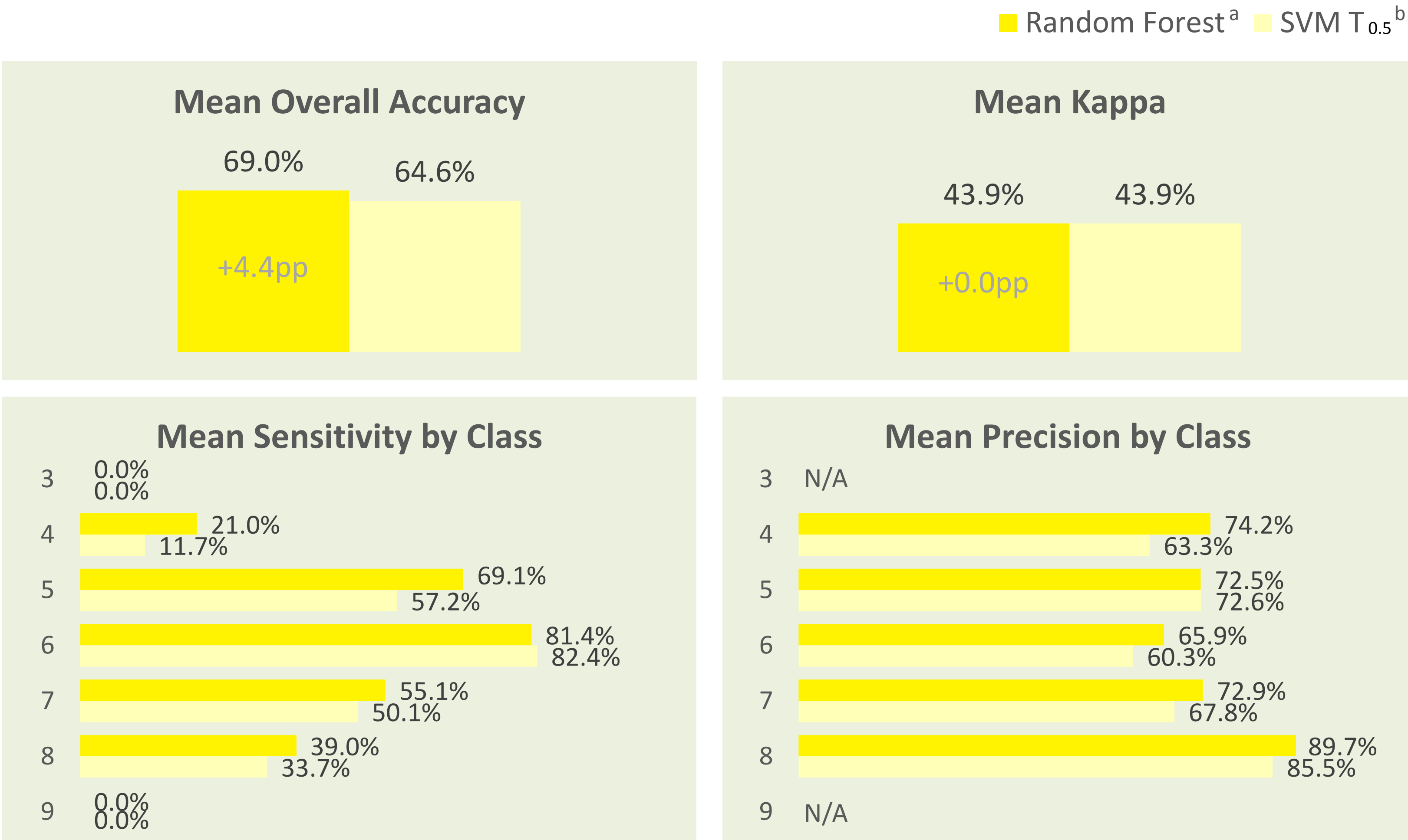



[a] 20 iterations of Scikit-Learn's Random Forest Classifier with stratified 5-fold cross validation with 500 estimators
[b] Cortez et. al.  Note, Sensitivity by Class calculated from published confusion matrix.

n = 1,599

# White: RF slightly better

## No lift for highest (9), lowest (3) classes



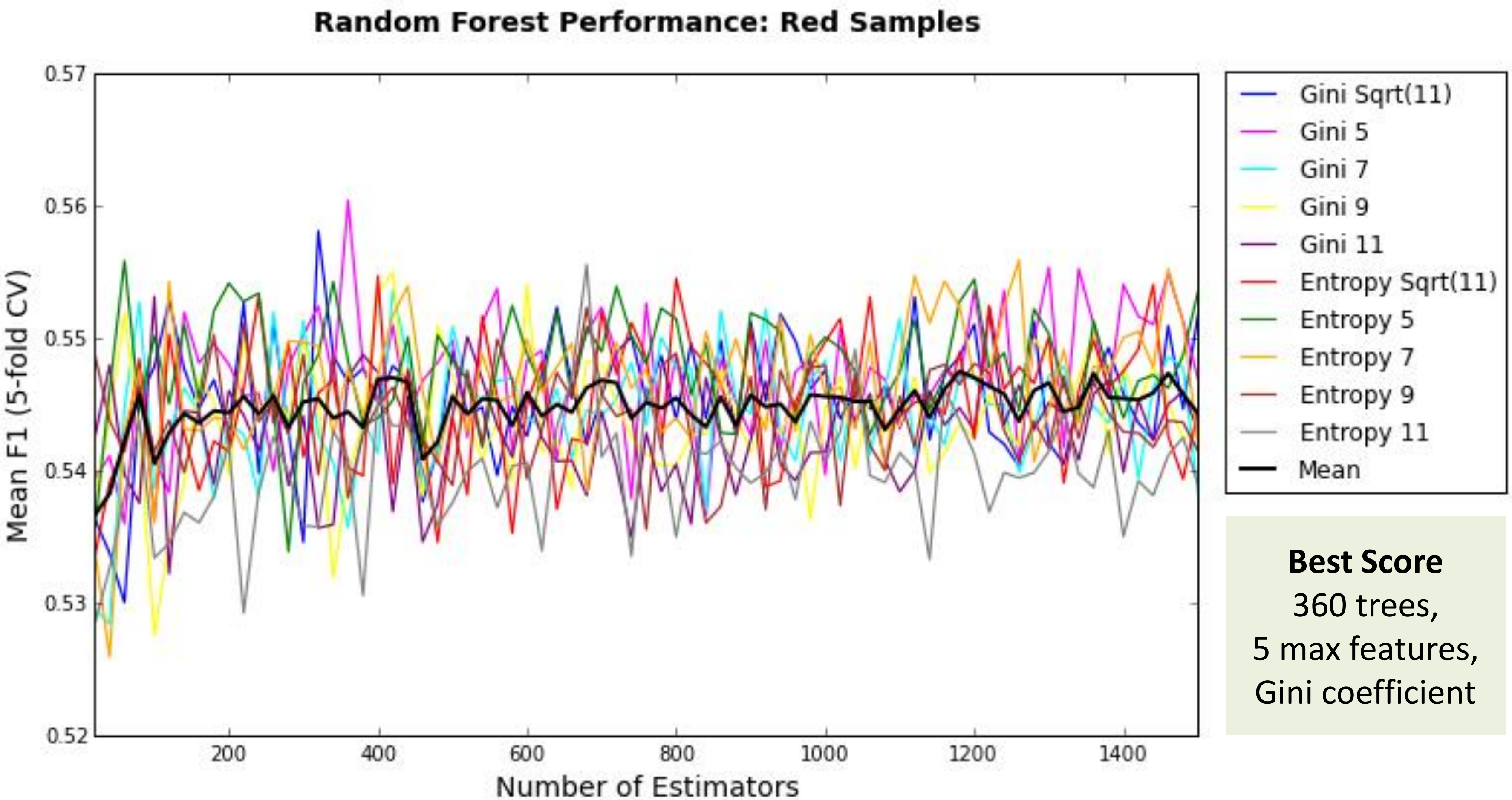

[a] 20 iterations of Scikit-Learn's Random Forest Classifier with stratified 5-fold cross validation with 500 estimators

[b] Cortez et. al. Note, Sensitivity by Class calculated from published confusion matrix.

n = 4,898

# Red: Tuning suggests 5 features, fewer trees

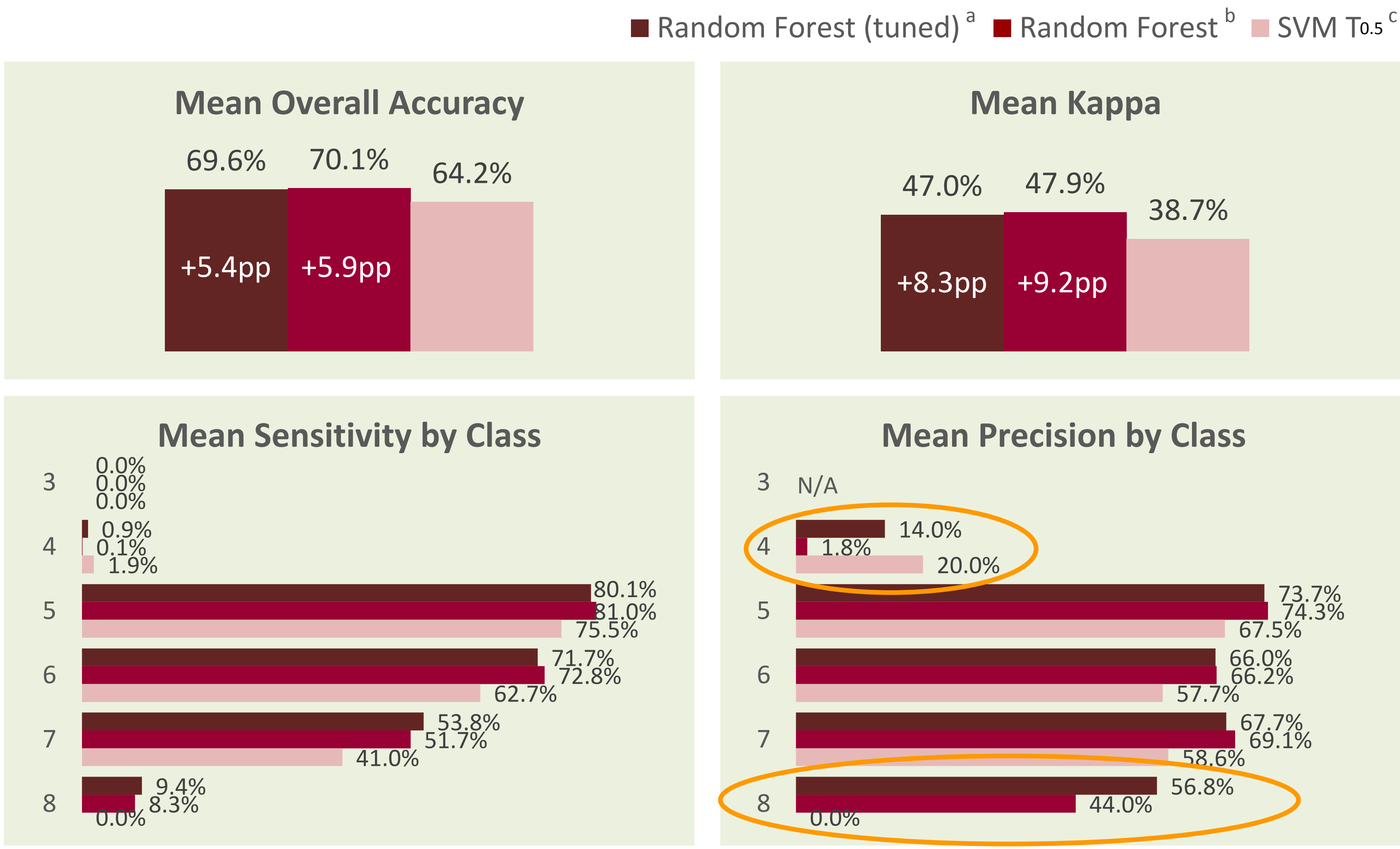
Improvement likely to be small, if any




**Best Score**
360 trees,
5 max features,
Gini coefficient

n = 1,599

# Red: Only hi/lo precision improves with tuning




[a] 20 iterations of Scikit-Learn's Random Forest Classifier with stratified 5-fold cross validation, 360 estimators and 5 max features

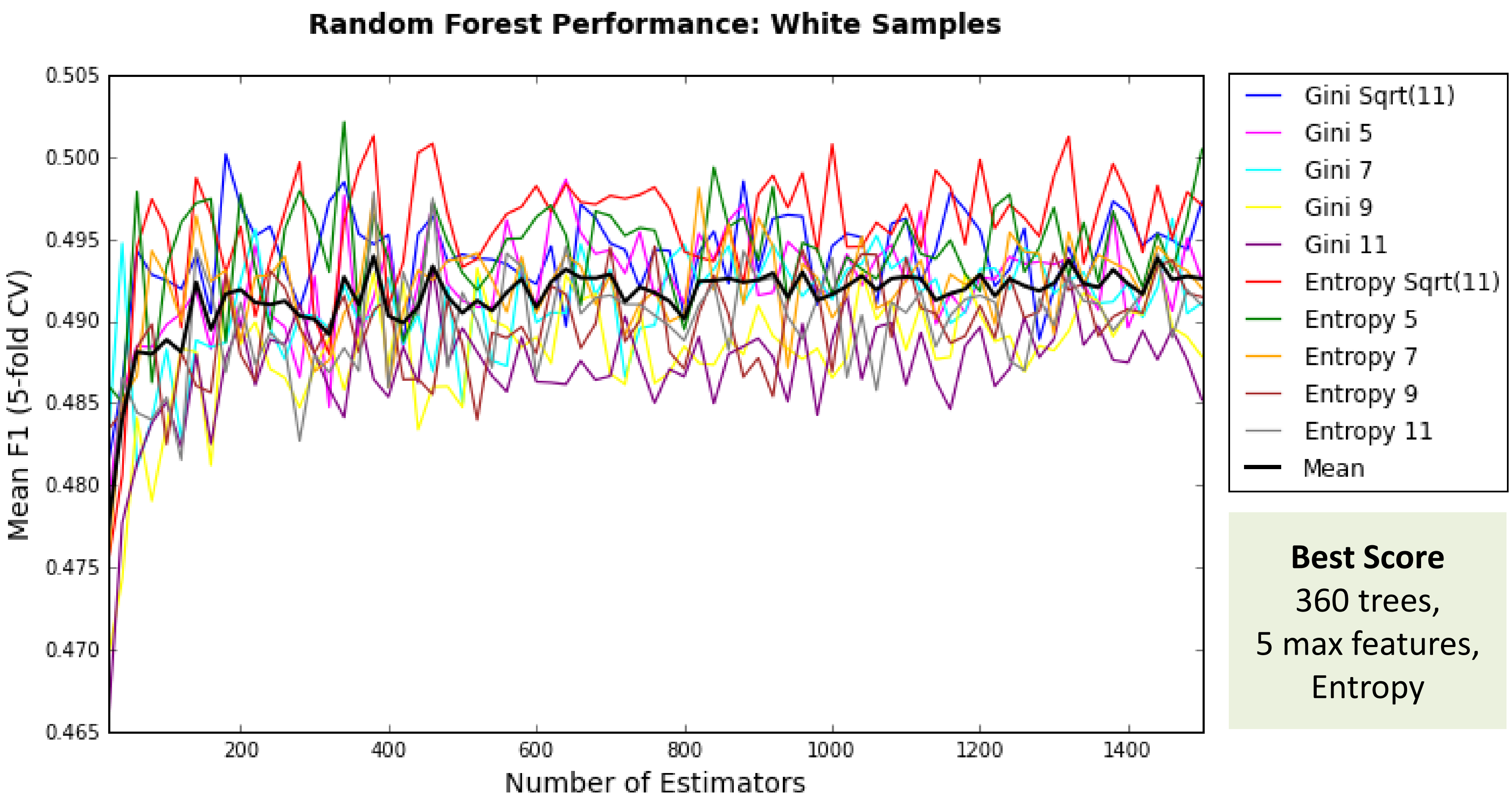
[b] 20 iterations of Scikit-Learn's Random Forest Classifier with stratified 5-fold cross validation and 500 estimators

[c] Cortez et. al. Note, Sensitivity by Class calculated from published confusion matrix.

n = 1,599

# White: Similar changes + entropy split criterion

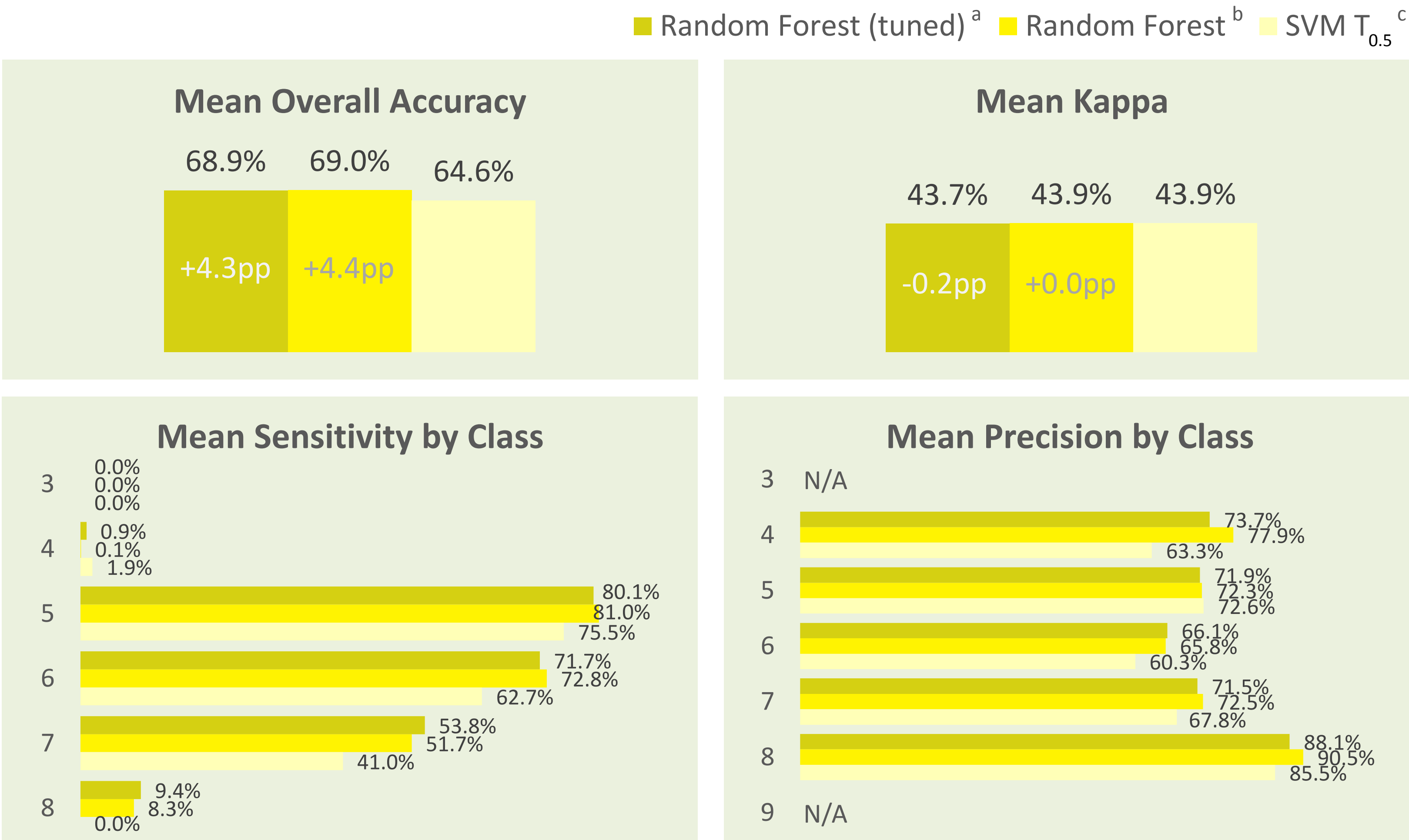
## Toss up between sqrt(d) and 5 max features




**Best Score**
360 trees,
5 max features,
Entropy

n = 4,898

# White: No notable change with tuning




[a] 20 iterations of Scikit-Learn's Random Forest Classifier with stratified 5-fold cross validation, 360 estimators, 5 max features and entropy

[b] 20 iterations of Scikit-Learn's Random Forest Classifier with stratified 5-fold cross validation and 500 estimators

[c] Cortez et. al. Note, Sensitivity by Class calculated from published confusion matrix.

n = 4,898

# Conclusions

- Random Forest appears to beat SVM, especially for red, but still room for improvement
  - Variable transformations? e.g., log
  - Other methods? e.g., Sklearn Multiclass
- Results potentially useful for wine production and marketing decisions
- Key challenge will be expanding dataset to cover other varietals, regions, metrics and/or audiences