Predicting Perceived *Vinho Verde* Quality using Random Forests

Megan McGoldrick
GA Data Science
April 2015

Objective

In 2009, a research team in Portugal published *Modeling Wine Preferences by Data Mining from Physiochemical Properties*,[1] which evaluated several modeling methods "to predict human wine taste preferences based on easily available analytical tests." Using perceived quality ratings from blind taste tests and measures of physiochemical properties of red and white *Vinho Verde* samples, they demonstrated that the support vector machine (SVM) method outperformed neural network (NN) and multiple regression methods on several measures of classification performance.

My goal is to show whether or not an alternative machine learning method can produce comparable or better classification performance of perceived *Vinho Verde* quality with less effort spent on data transformation, parameter tuning and/or feature selection. My hypothesis is that the Random Forests ensemble method is well suited to this classification problem for reasons later discussed and is the focus of my analysis.

Dataset

The data for my analysis are available through the UCI Machine Learning Repository[2] . They were original collected during routine laboratory testing of *Vinho Verde* wine samples by the CVRVV, an inter-professional organization that serves as the official wine certification body in Portugal, and shared by the authors of following their published analysis.

The data comprise two sets: one with 1,599 samples of red and one with 4,898 samples of white *Vinho Verde* from northwest Portugal, tested between May 2004 and February 2007. Each set contains 12 variables, 11 of which are numeric measurements of various physiochemical properties (see figure 1 for descriptions) and one of which is a measure of sensory-based, perceived quality for each sample. The quality score is the median of ratings on a 0 to 10 (poor to excellent) scale from three or more experts who participated in blind tastings. Quality will serve as the classification outcome and the physiochemical properties as explanatory features. There are no missing values.

[1] 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.
[2] http://archive.ics.uci.edu/ml/datasets/Wine+Quality

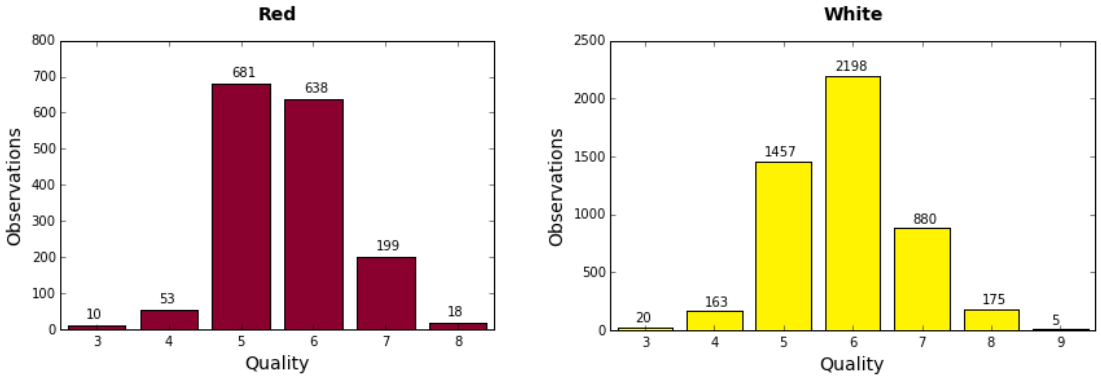
Figure 1. Explanatory Feature Descriptions

Physiochemical Property	Unit	Description
Fixed Acidity	g/dm^3	Naturally occurring tartaric acid
Volatile Acidity	g/dm^3	Acetic acid produced during fermentation that leads to unpleasant aromas
Citric Acid	g/dm^3	Inexpensive supplement to boost total acidity
pH		Inversely related to acidity levels; higher levels make wines taste softer
Alcohol	% volume	Produced from yeast and sugar during fermentation; higher levels associated with fuller-bodied, more complex wines
Residual Sugar	g/dm^3	Remaining sugar after fermentation or from added, unfermented must; higher levels yield sweeter wines
Density	g/dm^3	Similar to the density of water; lower for "dryer" wines, higher for sweeter wines
Chlorides	g/dm^3	“Saltiness,” often just barely detectable
Sulphates	g/dm^3	Potassium sulphate used to lower pH, raise acidity and intensify color in red wines
Free Sulfur Dioxide Total Sulfur Dioxide	g/dm^3 g/dm^3	Serve as antibiotics/antioxidants to protect wine and used in winery sanitation; detectable as pungent odor at higher levels and responsible for label “contains sulfites”

Exploratory Analysis

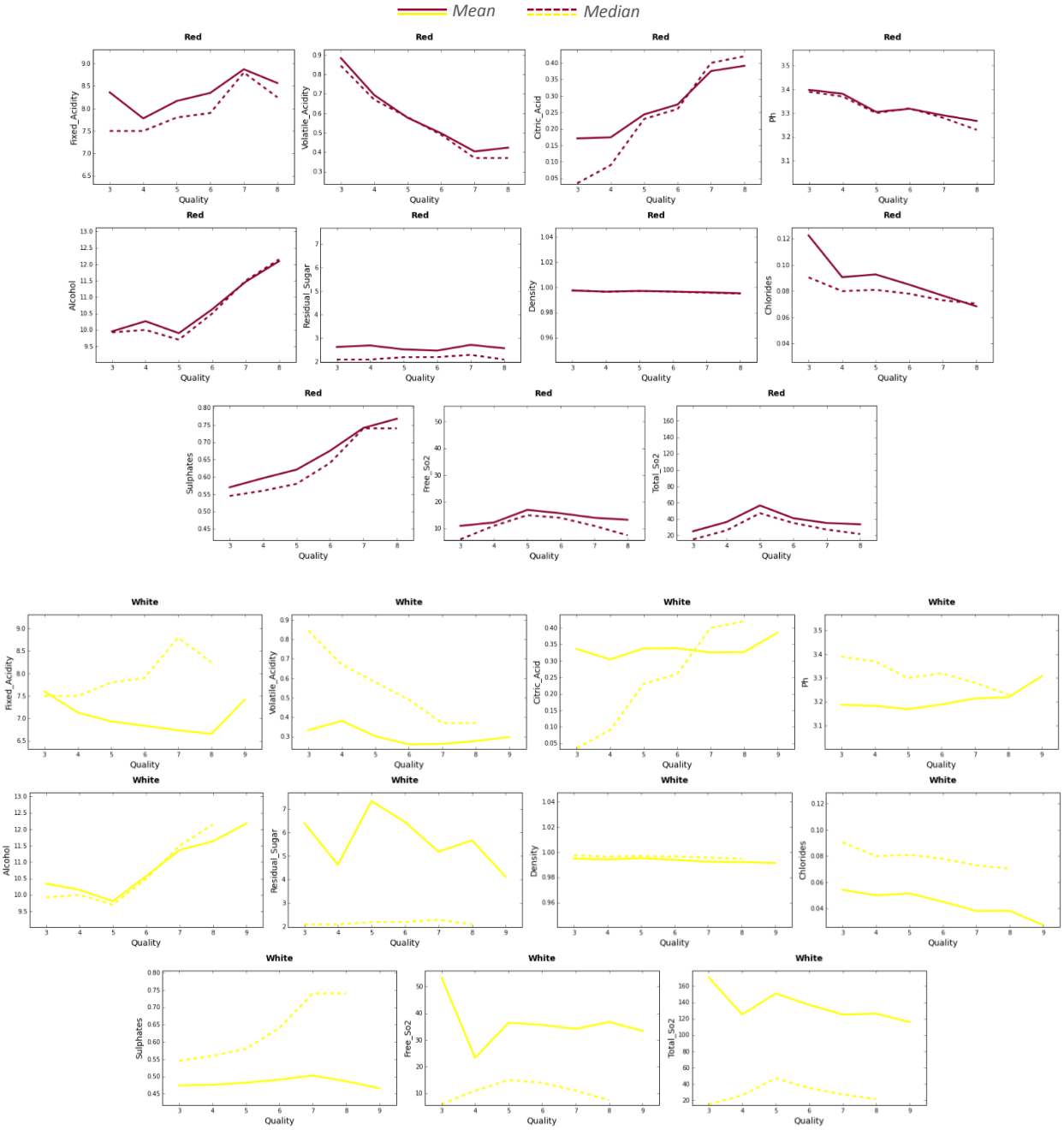
Exploratory analysis of the data reveals several challenges with this classification problem. First, the quality outcome has multiple, imbalanced classes (figures 2-3). The rating scale ranges from zero to 10, although most *Vinho Verde* samples are perceived as “average:” 82% of red samples and 75% of white samples have a median quality rating of 5 or 6. Relatively few samples have a median quality rating of 3, 4, 7, 8 or 9, and none have a median rating of 0, 1, 2 or 10.

Figure 2-3. Distribution of Quality Scores (Classes)



Second, plots of feature means and medians by quality (figure4-25) show that some explanatory features, especially in the white dataset, are skewed. Also, the relationships between quality and several explanatory features appear to are non-linear and differ among red and white samples.

Figures 4-25. Explanatory Features versus Quality Scores (Classes)



Lastly, several explanatory features are dependent on one another, evidenced by strong Spearman Rank correlation coefficients (figure 26). These dependencies are inherent to the nature of wine and the wine making process. For example, during fermentation, yeast converts sugar in grapes to alcohol; as alcohol rises, residual sugar and density decline. Similarly, pH will be higher in wines with lower acidity.

Figure 26. Spearman Rank Correlation Coefficients

Red White

	Fixed Acidity	Volatile Acidity	Citric Acid	Residual Sugar	Chlorides	Free S02	Total SO2	Density	pH	Sulphates	Alcohol
Fixed Acidity		-0.3	0.7	0.2	0.3	-0.2	-0.1	0.6	-0.7	0.2	-0.1
Volatile Acidity	0.0		-0.6	0.0	0.2	0.0	0.1	0.0	0.2	-0.3	-0.2
Citric Acid	0.3	-0.2		0.2	0.1	-0.1	0.0	0.4	-0.5	0.3	0.1
Residual Sugar	0.1	0.1	0.0		0.2	0.1	0.1	0.4	-0.1	0.0	0.1
Chlorides	0.1	0.0	0.0	0.2		0.0	0.1	0.4	-0.2	0.0	-0.3
Free S02	0.0	-0.1	0.1	0.3	0.2		0.8	0.0	0.1	0.0	-0.1
Total SO2	0.1	0.1	0.1	0.4	0.4	0.6		0.1	0.0	0.0	-0.3
Density	0.3	0.0	0.1	0.8	0.5	0.3	0.6		-0.3	0.2	-0.5
pH	-0.4	0.0	-0.1	-0.2	-0.1	0.0	0.0	-0.1		-0.1	0.2
Sulphates	0.0	0.0	0.1	0.0	0.1	0.1	0.2	0.1	0.1		0.2
Alcohol	-0.1	0.0	0.0	-0.4	-0.6	-0.3	-0.5	-0.8	0.1	0.0	1.0

Modeling Approach

Given the objective of my analysis and the challenges just noted with the data, the Random Forest ensemble method is likely the best choice for predicting perceived wine quality. This method does not require feature scaling or much tuning up front, works well with non-linear data, and minimizes the impact of outliers and irrelevant features through fuller coverage of the dataset.

A summary of my approach is as follows.

1. Perform 20 iterations of Scikit-Learn's Random Forest Classifier with 500 trees and stratified 5-fold cross validation on the full dataset. Compare results to those published for SVM.
2. Perform grid search with F1 scoring and 5-fold cross validation on the full dataset to tune parameters for number of trees (20 to 1500), maximum features per tree (default sqrt(d) vs 5, 7, 9 and 11) and different measures of tree split quality (gini impurity vs entropy).
3. Perform 20 iterations of Scikit-Learn's Random Forest Classifier with tuned parameters for "best" model and stratified 5-fold cross validation on the full dataset. Compare results to those published for SVM.

To evaluate performance, I examine overall accuracy, kappa, sensitivity by class and precision by class for each Random Forest classifier – absolute and relative to the published results for SVM with a tolerance of 0.5.[3] It is important to note that there may be some differences between how the two sets of results are produced and calculated, but our approaches are likely close enough to justify benchmarking Random Forest against SVM.

There are several steps in my process that warrant further explanation. First, although Random Forests by design are an iterative, "averaging" method, I choose to perform 20 iterations of each classifier in order to

[3] The authors report classification results for three rounding tolerance thresholds: 0.25, 0.5 and 1.0. It is unclear whether the rounding occurs before or after the 20 model iterations are averaged.

compute 95% confidence intervals for overall accuracy and kappa. These intervals demonstrate how stable the classifier is and how much better or worse Random Forest really performs relative to SVM. Second, I use 5-fold cross validation to mimic the published approach to SVM and because the number of folds must be greater than or equal to the smallest class; in this case, n=5 for white class 9. Third, I select F1 as my scoring criterion for tuning, as this measure is more sensitive to smaller classes. Lastly, I perform my analysis separately for red and white samples, as this mimics the published approach to SVM and seems justified by the differences between the samples on a number of explanatory features.

Results

Figure 27 shows the mean results over 20 iterations for my original and tuned Random Forest classifiers, as well as for the published SVM classifier with a tolerance of 0.5. Figures 28-29 illustrate the results from parameter tuning with grid search for the Random Forest classifiers.

Figure 27. Classifier Performance

Mean % over 20 iterations	Red			White		
	Random Forest (initial)	**Random Forest (tuned)**	**Published SVM $T_{0.5}$**	**Random Forest (initial)**	**Random Forest (tuned)**	**Published SVM $T_{0.5}$**
Overall Accuracy	70.1 ± 0.3	69.6 ± 0.3	64.2± 0.4	69.0 ± 0.2	68.9 ± 0.2	64.6 ± 0.4
Kappa	47.9 ± 0.6	47.0 ± 0.6	38.7 ± 0.7	43.9 ± 0.4	43.7 ± 0.3	43.9 ± 0.4
Sensitivity						
3	0.0	0.0	0.0	0.0	0.0	0.0
4	0.1	0.9	1.9	21.0	22.8	11.7
5	81.0	80.1	75.5	69.1	69.4	57.2
6	72.8	71.7	62.7	81.4	80.6	82.4
7	51.7	53.8	41.0	55.1	55.4	50.1
8	8.3	9.4	0.0	39.0	40.2	33.7
9	N/A	N/A	N/A	0.0	0.0	0.0
Precision						
3	N/A	N/A	N/A	N/A	N/A	N/A
4	1.8	14.0	20.0	77.9	73.7	63.3
5	74.3	73.7	67.5	72.3	71.9	72.6
6	66.2	66.0	57.7	65.8	66.1	60.3
7	69.1	67.7	58.6	72.5	71.5	67.8
8	44.0	56.8	0.0	90.5	88.1	85.5
9	N/A	N/A	N/A	N/A	N/A	N/A

Red Performance

The initial Random Forest classifier for red samples achieves a mean overall accuracy of70.1% with a 95% confidence interval of ±0.3% and a Cohen's kappa of 47.9% with a 95% confidence interval of ±0.6%. This classifier outperforms the published SVM $T_{0.5}$ mean results by 5.9 and 9.2 percentage points, respectively. The initial Random Forest classifier also has higher sensitivity and precision for classes 5 to 8, while SVM is better on class 4. Neither does well predicting the lowest class 3.

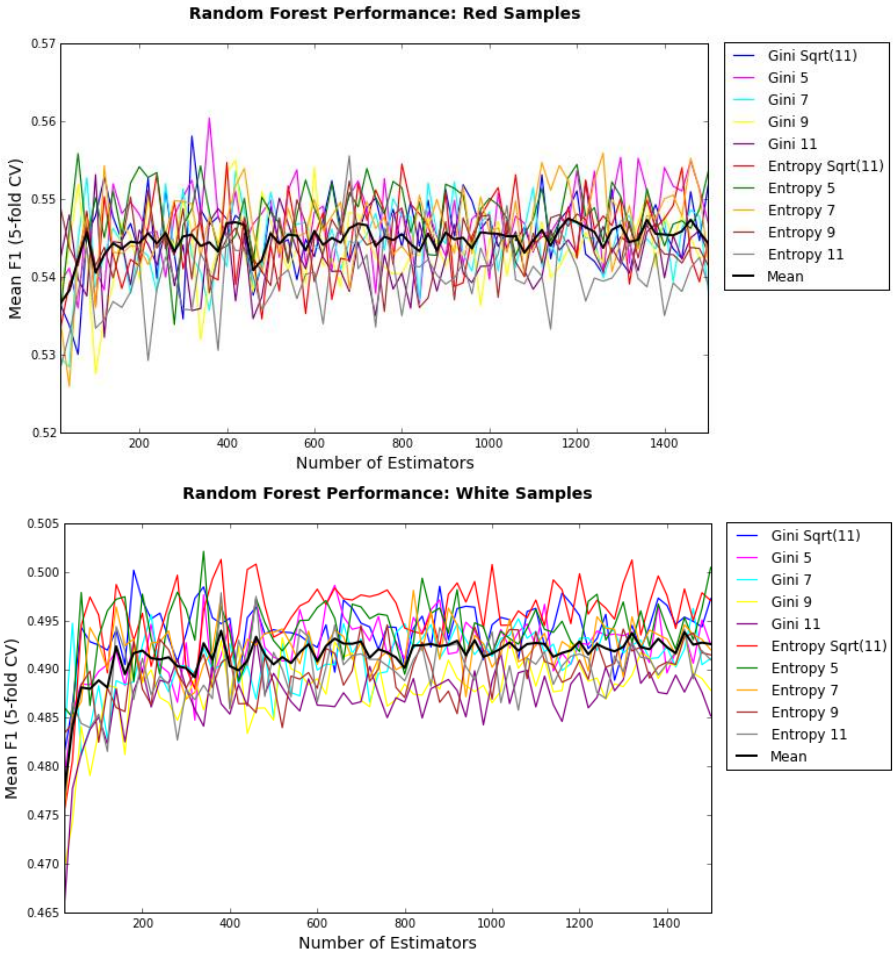
Grid search suggests that the "best" Random Forest classifier for red samples, using F1 scoring, has 360 trees instead of 500 and a maximum of five features instead of sqrt (11); gini impurity is still considered the best measure of tree split quality. However, over 20 iterations, the tuned Random Forest classifier does not perform any better than the initial classifier, except on class 7 sensitivity and class 4 and 8 precision.

White Performance

The initial Random Forest classifier for white samples achieves a mean overall accuracy of 69.0% with a 95% confidence interval of ±0.2% and a Cohen's kappa of 43.7% with a 95% confidence interval of ±0.4%. This classifier achieves slightly better mean overall accuracy (+ 4.4 percentage points) and the same mean kappa as the published SVM $T_{0.5}$ results. On sensitivity and precision by class, the initial Random Forest classifier is better on classes 4, 7 and 8, and comparable on classes 5, 6. Neither does well predicting the lowest (3) or highest (9) classes.

Grid search suggests that the "best" Random Forest classifier for white samples, using F1 scoring, has 360 trees instead of 500, a maximum of five features instead of sqrt (11), and entropy instead of gini impurity as the measure of tree split quality. However, over 20 iterations, the tuned Random Forest classifier does not perform any better than the initial classifier.

Figure 28-29. Random Forest Parameter Tuning Results



Feature Importance

Figures 30-31 shows the mean feature importance scores over 20 iterations for the initial (untuned) red and white Random Forest classifiers and their rank order versus that published for SVM $T0_{.5}$. All features appear to be important, with alcohol topping the list with both red and white Random Forest classifiers. However, there are differences in rank order across wine types and classification methods.

Figures 30-31. Feature Importance

Red	RF Importance	RF Rank	SVM $T_{0.5}$ Rank
Alcohol	0.146	1	4
Sulphates	0.111	2	1
Total SO2	0.105	3	3
Volatile Acidity	0.104	4	5
Density	0.092	5	10
Chlorides	0.080	6	9
Fixed Acidity	0.075	7	7
pH	0.075	8	2
Citric Acid	0.074	9	11
Residual Sugar	0.071	10	8
Free SO2	0.067	11	6

White	RF Importance	RF Rank	SVM $T_{0.5}$ Rank
Alcohol	0.115	1	2
Density	0.104	2	8
Volatile Acidity	0.099	3	7
Free SO2	0.094	4	6
Total SO2	0.092	5	5
Residual Sugar	0.088	6	3
pH	0.086	7	9
Chlorides	0.085	8	10
Citric Acid	0.081	9	4
Sulphates	0.080	10	1
Fixed Acidity	0.075	11	11

Conclusion

Random Forests appear to offer some performance advantages over the published SVM approach for predicting *Vinho Verde* quality, especially for red samples. However, overall accuracy is only about 70%, and predictions for the lowest (3) and highest (9) classes are still unattainable. With additional time, I would like to explore alternative methods, such as Scikit-learn's Multiclass, and potential useful data transformations, to try to improve results.

I believe this type of analysis has the potential to impact important wine making and marketing decisions. For example, estimating perceived quality earlier in the wine making process may enable winemakers to modify decisions, such as adding more or less citric acid or sulphates. Quality predictions after a wine is produced can also help teams to adjust labeling or pricing to align with expected quality. To bring these possibilities to fruition, the dataset would need to be expanded to cover other varietals and regions. I would also like to capture a variety of other variables, such as grape growing conditions (soil composition, weather) and fermentation and aging practices.

Appendix: Feature Descriptive Statistics

RED	Fixed Acidity	Volatile Acidity	Citric Acid	Residual Sugar	Chlorides	Free S02	Total SO2	Density	pH	Sulphates	Alcohol	Quality
Count	1599	1599	1599	1599	1599	1599	1599	1599	1599	1599	1599	1599
Mean	8.32	0.53	0.27	2.54	0.09	15.87	46.47	1.00	3.31	0.66	10.42	5.64
Std	1.74	0.18	0.19	1.41	0.05	10.46	32.90	0.00	0.15	0.17	1.07	0.81
Min	4.6	0.12	0	0.9	0.012	1	6	0.990	2.74	0.33	8.4	3
25%	7.1	0.39	0.09	1.9	0.07	7	22	0.996	3.21	0.55	9.5	5
50%	7.9	0.52	0.26	2.2	0.079	14	38	0.997	3.31	0.62	10.2	6
75%	9.2	0.64	0.42	2.6	0.09	21	62	0.998	3.4	0.73	11.1	6
Max	15.9	1.58	1	15.5	0.611	72	289	1.004	4.01	2	14.9	8

WHITE	Fixed Acidity	Volatile Acidity	Citric Acid	Residual Sugar	Chlorides	Free S02	Total SO2	Density	pH	Sulphates	Alcohol	Quality
Count	4898	4898	4898	4898	4898	4898	4898	4898	4898	4898	4898	4898
Mean	6.85	0.28	0.33	6.39	0.05	35.31	138.36	0.99	3.19	0.49	10.51	5.88
Std	0.84	0.10	0.12	5.07	0.02	17.01	42.50	0.00	0.15	0.11	1.23	0.89
Min	3.8	0.08	0	0.6	0.009	2	9	0.987	2.72	0.22	8	3
25%	6.3	0.21	0.27	1.7	0.036	23	108	0.992	3.09	0.41	9.5	5
50%	6.8	0.26	0.32	5.2	0.043	34	134	0.994	3.18	0.47	10.4	6
75%	7.3	0.32	0.39	9.9	0.05	46	167	0.996	3.28	0.55	11.4	6
Max	14.2	1.1	1.66	65.8	35%	289	440	1.039	3.82	108%	14.2	9