“Vinho Verde” Wines Quality Modeling

The First Group (T.F.G.)

August 14, 2018

Can Robots Taste Wine?

Vihno Verde - 2000 Years of Winemaking

Wines Dataset Attributes

```
6497 observations:

Input variables (based on physicochemical tests):
  1 - fixed acidity        (FA)
  2 - volatile acidity     (VA)
  3 - citric acid          (CA)
  4 - residual sugar       (RS)
  5 - chlorides            (CH)
  6 - free sulfur dioxide  (FSD)
  7 - total sulfur dioxide (TSD)
  8 - density              (DEN)
  9 - pH                   (pH)
  10 - sulphates           (SUL)
  11 - alcohol             (ALC)

Output variable (based on sensory data):
  12 - quality (score between 0 and 10) - (QLT)
```

Wines Quality Dataset - First Rows

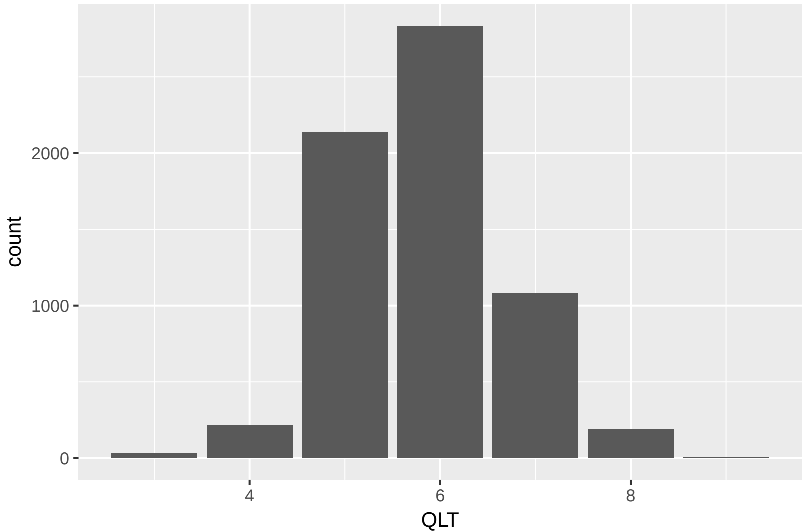
	FA	VA	CA	RS	CH	FSD	TSD	DEN	pH	SUL	ALC	QLT	TYPE
1	7.40	0.70	0.00	1.90	0.08	11.00	34.00	1.00	3.51	0.56	9.40	5	0.00
2	7.80	0.88	0.00	2.60	0.10	25.00	67.00	1.00	3.20	0.68	9.80	5	0.00
3	7.80	0.76	0.04	2.30	0.09	15.00	54.00	1.00	3.26	0.65	9.80	5	0.00
4	11.20	0.28	0.56	1.90	0.07	17.00	60.00	1.00	3.16	0.58	9.80	6	0.00
5	7.40	0.70	0.00	1.90	0.08	11.00	34.00	1.00	3.51	0.56	9.40	5	0.00
6	7.40	0.66	0.00	1.80	0.07	13.00	40.00	1.00	3.51	0.56	9.40	5	0.00
7	7.90	0.60	0.06	1.60	0.07	15.00	59.00	1.00	3.30	0.46	9.40	5	0.00
8	7.30	0.65	0.00	1.20	0.06	15.00	21.00	0.99	3.39	0.47	10.00	7	0.00
9	7.80	0.58	0.02	2.00	0.07	9.00	18.00	1.00	3.36	0.57	9.50	7	0.00
10	7.50	0.50	0.36	6.10	0.07	17.00	102.00	1.00	3.35	0.80	10.50	5	0.00
11	6.70	0.58	0.08	1.80	0.10	15.00	65.00	1.00	3.28	0.54	9.20	5	0.00
12	7.50	0.50	0.36	6.10	0.07	17.00	102.00	1.00	3.35	0.80	10.50	5	0.00
13	5.60	0.61	0.00	1.60	0.09	16.00	59.00	0.99	3.58	0.52	9.90	5	0.00
14	7.80	0.61	0.29	1.60	0.11	9.00	29.00	1.00	3.26	1.56	9.10	5	0.00
15	8.90	0.62	0.18	3.80	0.18	52.00	145.00	1.00	3.16	0.88	9.20	5	0.00
16	8.90	0.62	0.19	3.90	0.17	51.00	148.00	1.00	3.17	0.93	9.20	5	0.00
17	8.50	0.28	0.56	1.80	0.09	35.00	103.00	1.00	3.30	0.75	10.50	7	0.00
18	8.10	0.56	0.28	1.70	0.37	16.00	56.00	1.00	3.11	1.28	9.30	5	0.00
19	7.40	0.59	0.08	4.40	0.09	6.00	29.00	1.00	3.38	0.50	9.00	4	0.00
20	7.90	0.32	0.51	1.80	0.34	17.00	56.00	1.00	3.04	1.08	9.20	6	0.00

Dataset Attributes Summary

FA	VA	CA	RS	CH	FSD	TSD
Min. : 3.800	Min. :0.0800	Min. :0.0000	Min. : 0.600	Min. :0.00900	Min. : 1.00	Min. : 6.0
1st Qu.: 6.400	1st Qu.:0.2300	1st Qu.:0.2500	1st Qu.: 1.800	1st Qu.:0.03800	1st Qu.: 17.00	1st Qu.: 77.0
Median : 7.000	Median :0.2900	Median :0.3100	Median : 3.000	Median :0.04700	Median : 29.00	Median :118.0
Mean : 7.215	Mean :0.3397	Mean :0.3186	Mean : 5.443	Mean :0.05603	Mean : 30.53	Mean :115.7
3rd Qu.: 7.700	3rd Qu.:0.4000	3rd Qu.:0.3900	3rd Qu.: 8.100	3rd Qu.:0.06500	3rd Qu.: 41.00	3rd Qu.:156.0
Max. :15.900	Max. :1.5800	Max. :1.6600	Max. :65.800	Max. :0.61100	Max. :289.00	Max. :440.0

TSD	DEN	pH	SUL	ALC	QLT	TYPE
Min. : 6.0	Min. :0.9871	Min. :2.720	Min. :0.2200	Min. : 8.00	Min. :3.000	Min. :0.0000
1st Qu.: 77.0	1st Qu.:0.9923	1st Qu.:3.110	1st Qu.:0.4300	1st Qu.: 9.50	1st Qu.:5.000	1st Qu.:1.0000
Median :118.0	Median :0.9949	Median :3.210	Median :0.5100	Median :10.30	Median :6.000	Median :1.0000
Mean :115.7	Mean :0.9947	Mean :3.219	Mean :0.5313	Mean :10.49	Mean :5.818	Mean :0.7539
3rd Qu.:156.0	3rd Qu.:0.9970	3rd Qu.:3.320	3rd Qu.:0.6000	3rd Qu.:11.30	3rd Qu.:6.000	3rd Qu.:1.0000
Max. :440.0	Max. :1.0390	Max. :4.010	Max. :2.0000	Max. :14.90	Max. :9.000	Max. :1.0000

Distribution of QLT in the Dataset



Random Forest Regressor Modeling

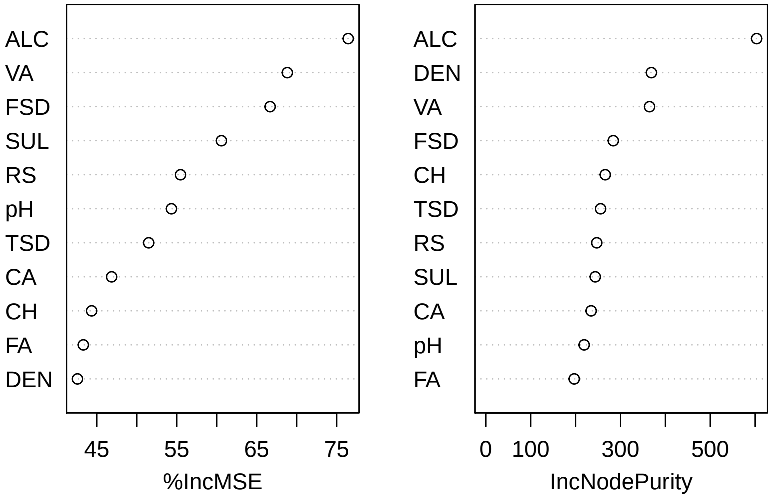
```
library(randomForest)
fitRF1 <- randomForest(
  QLT ~ ., method="anova",
  data=train1.data, importance=TRUE, ntree=500)

PredictionRF1 <- predict(fitRF1, test1.data)

cor(PredictionRF1,test1.data$QLT)
```

```
## [1] 0.7114132
```

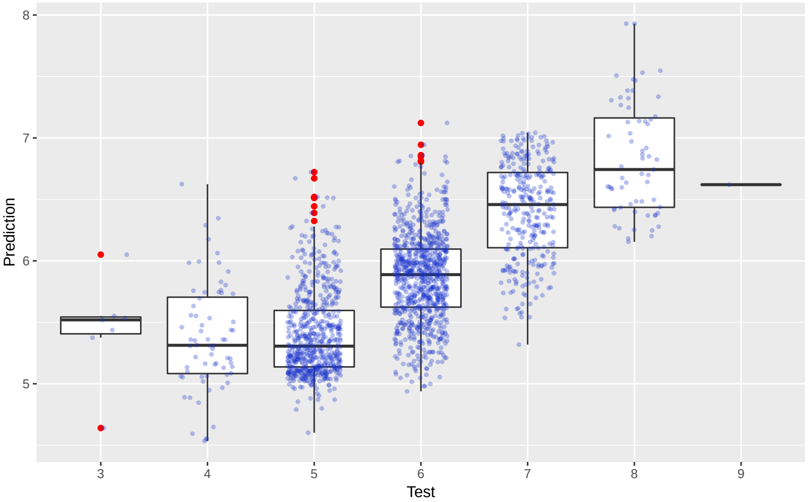
Importance of the Dataset Attributes for QLT Prediction



Random Forest Pledictor Confusion Matrix

	3	4	5	6	7	8	9
5	3	45	441	145	1	0	0
6	4	22	194	670	168	21	0
7	0	1	5	35	150	37	1
8	0	0	0	0	0	5	0

Random Forest Prediction Scatter Plot



SVM Modeling and Accuracy

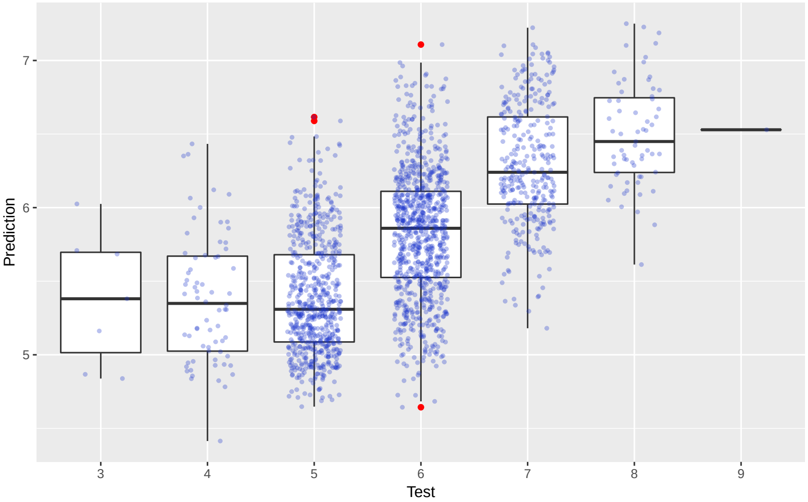
```
library("e1071")
svm_model <- svm(QLT ~ ., data=train1.data)
predSVM <- predict(svm_model, test1.data)
cor(predSVM,test1.data$QLT)
```

```
## [1] 0.6132799
```

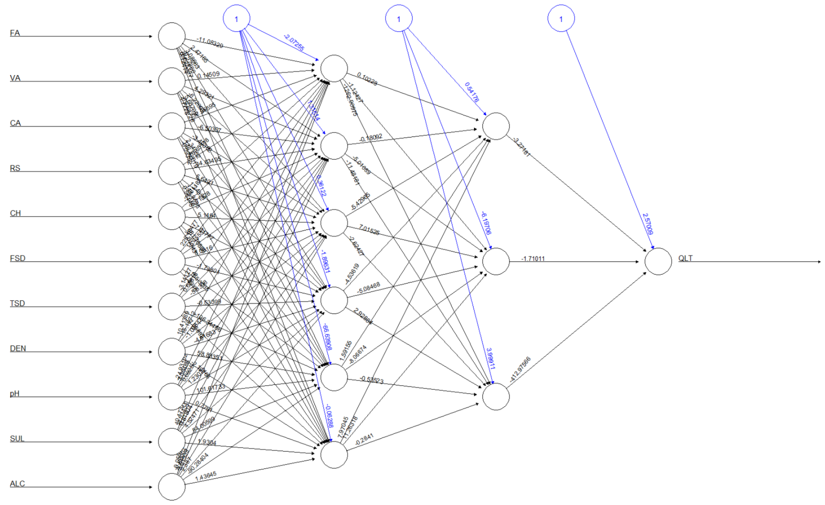
SVM Pledictor Confusion Matrix

	3	4	5	6	7	8	9
4	0	1	0	0	0	0	0
5	4	43	418	203	9	0	0
6	3	24	220	589	209	32	0
7	0	0	2	58	101	31	1

SVM Prediction Scatter Plot



Neural Network Modeling

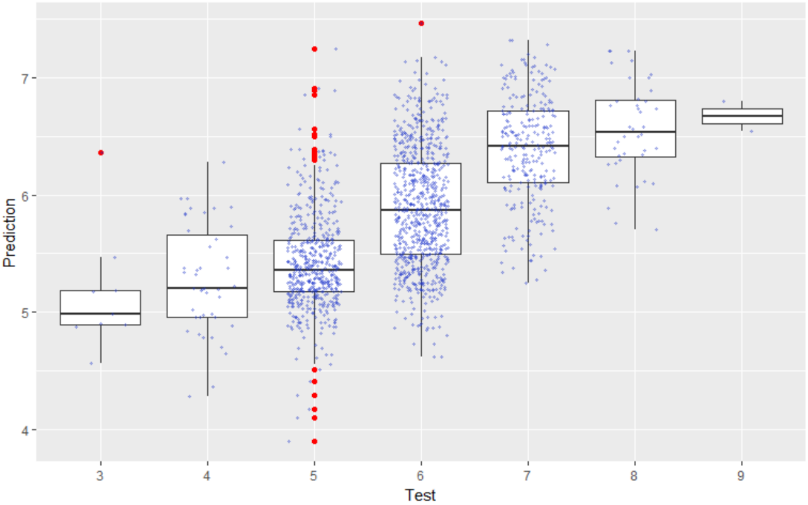


Neural Networks Pledictor Confusion Matrix

	3	4	5	6	7	8	9
4	0	2	5	0	0	0	0
5	8	28	336	189	10	0	0
6	1	13	179	456	137	19	0
7	0	0	6	97	115	21	2

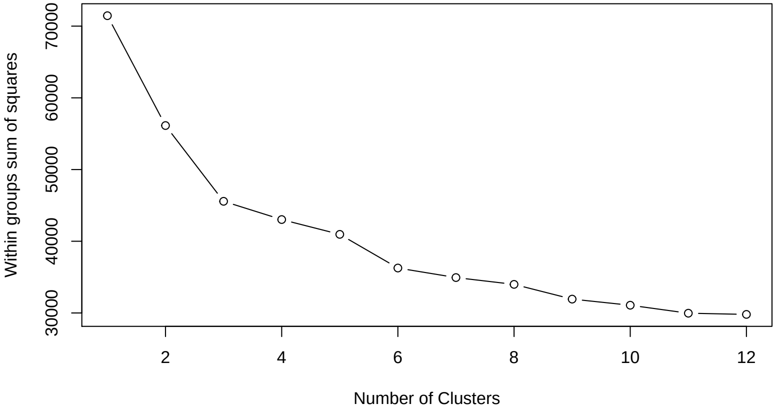
```
## [1] 0.6043741164
```

Neural Network Prediction Scatter Plot



Can we Guess Wine Type by its Biochemical Content?

Cluster Analysis K-Means - ‘Elbow Criterion’



Cluster Analysis K-Means - Cluster Centers

	FA	VA	CA	RS	CH	FSD	TSD	DEN	pH	SUL	ALC
1	-0.18	-0.35	0.28	1.20	-0.09	0.85	0.96	0.76	-0.39	-0.26	-0.80
2	-0.35	-0.40	-0.01	-0.44	-0.44	-0.09	0.03	-0.85	-0.04	-0.28	0.57
3	0.88	1.18	-0.32	-0.60	0.94	-0.84	-1.20	0.71	0.54	0.84	-0.13

Cluster Analysis K-Means - 2D Presentation

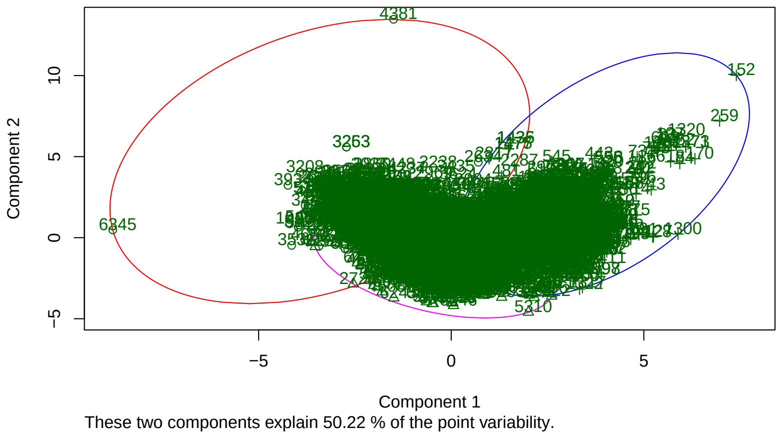


Figure 1: 2D representation of the Cluster solution

Cluster Analysis K-Means - no QLT Correlation

	1	2	3
3	12	8	10
4	48	100	68
5	804	638	696
6	843	1371	622
7	157	740	182
8	30	148	15
9	1	4	0

Wine Type is Correlated to Clusters!

	1	2	3
Red Wine	4	57	1538
White Wine	1891	2952	55

What's the Difference Between White Wines 1 and 2?

	Difference
RS	1.64
DEN	1.61
ALC	-1.36
FSD	0.94
TSD	0.92
CH	0.35
pH	-0.35
CA	0.28
FA	0.17
VA	0.05
SUL	0.03

```
Cluster 1 - sweet white wines
Cluster 2 - dry white wines
```

More Wine Groups - Trying Second "Elbow" at 6

	FA	VA	CA	RS	CH	FSD	TSD	DEN	pH	SUL	ALC
1	-0.17	-0.35	0.32	1.46	-0.14	0.93	1.00	0.92	-0.50	-0.28	-0.88
2	-0.55	-0.26	-0.03	-0.47	-0.59	-0.11	-0.16	-1.34	0.01	-0.28	1.43
3	2.01	0.50	0.96	-0.56	1.27	-0.90	-1.25	0.99	-0.07	1.42	0.04
4	0.09	1.68	-1.25	-0.63	0.68	-0.80	-1.16	0.49	0.96	0.40	-0.24
5	-0.60	-0.51	-0.16	-0.27	-0.27	0.40	0.54	-0.29	0.76	0.03	-0.18
6	0.13	-0.48	0.26	-0.31	-0.23	-0.28	0.05	-0.48	-0.77	-0.49	-0.02

Cluster Analysis K-Means - More Wine Types

	1	2	3	4	5	6
Red Wine	2	39	624	901	22	11
White Wine	1479	1147	19	51	1020	1182

What's the Difference Between Red Wines 3 and 4?

	Difference
CA	2.21
FA	1.93
VA	-1.18
SUL	1.02
pH	-1.02
CH	0.59
DEN	0.49
ALC	0.28
FSD	-0.10
TSD	-0.09
RS	0.07

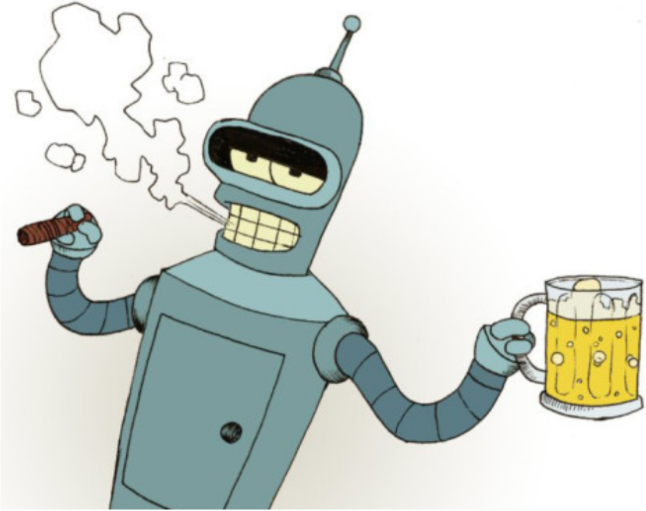
```
Cluster 3 - young fruity sour red wines
Cluster 4 - old red wines with a bit of bitterness
```

What Have We Learned?

- RF has the best overall prediction accuracy at 73%
- NN has better precision for best and worst wines
- CA can be used to recognize wine types and subtypes
- The Vihno Verde dataset missing some important attributes
- Robots can be the wine experts!

Enjoy Responsibly!

Questions?



Note

This slideshow was generated in R Studio using Beamer Knit template. The file itself is an executable R Markdown file that could be downloaded from Github with all the necessary artifacts:

https://github.com/ivbsoftware/big-data-final-2