“Vinho Verde” Wines Quality Modeling

The First Group (T.F.G.)

August 8, 2018

Wines Dataset Attributes

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

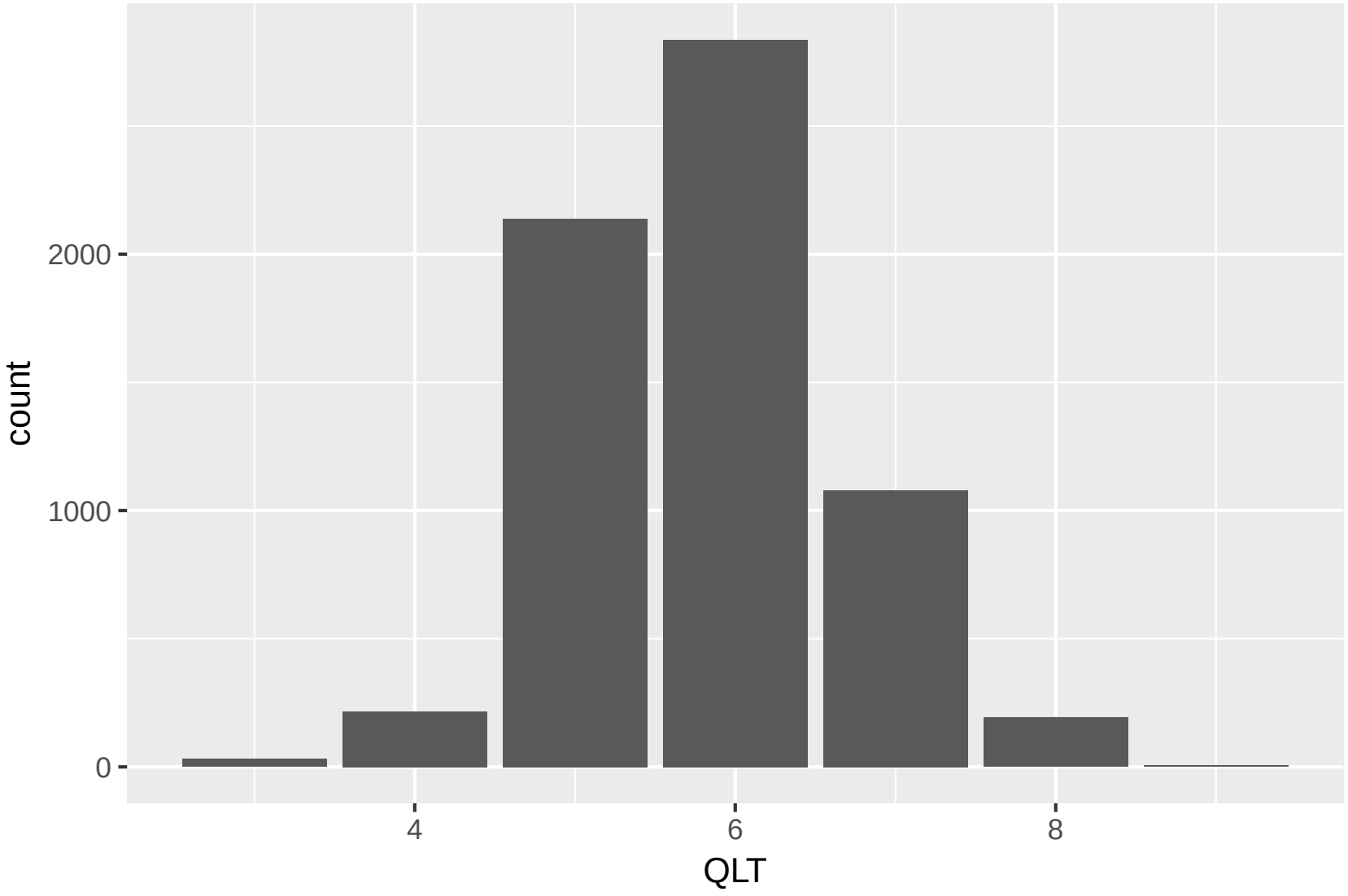
	FA	VA	CA	RS	CH	FSD	TSD	DEN	pH	SUL	ALC	QLT	TYPE
1	7.40	0.70	0.00	1.90	0.08	11.00	34.00	1.00	3.51	0.56	9.40	5	0.00
2	7.80	0.88	0.00	2.60	0.10	25.00	67.00	1.00	3.20	0.68	9.80	5	0.00
3	7.80	0.76	0.04	2.30	0.09	15.00	54.00	1.00	3.26	0.65	9.80	5	0.00
4	11.20	0.28	0.56	1.90	0.07	17.00	60.00	1.00	3.16	0.58	9.80	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
Min. : 3.800	Min. :0.0800	Min. :0.0000	Min. : 0.600	Min. :0.00900	Min. : 1.00
1st Qu.: 6.400	1st Qu.:0.2300	1st Qu.:0.2500	1st Qu.: 1.800	1st Qu.:0.03800	1st Qu.: 17.00
Median : 7.000	Median :0.2900	Median :0.3100	Median : 3.000	Median :0.04700	Median : 29.00
Mean : 7.215	Mean :0.3397	Mean :0.3186	Mean : 5.443	Mean :0.05603	Mean : 30.53
3rd Qu.: 7.700	3rd Qu.:0.4000	3rd Qu.:0.3900	3rd Qu.: 8.100	3rd Qu.:0.06500	3rd Qu.: 41.00
Max. :15.900	Max. :1.5800	Max. :1.6600	Max. :65.800	Max. :0.61100	Max. :289.00

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

Distribution of target value in the dataset



Random Forests Regressor Modeling

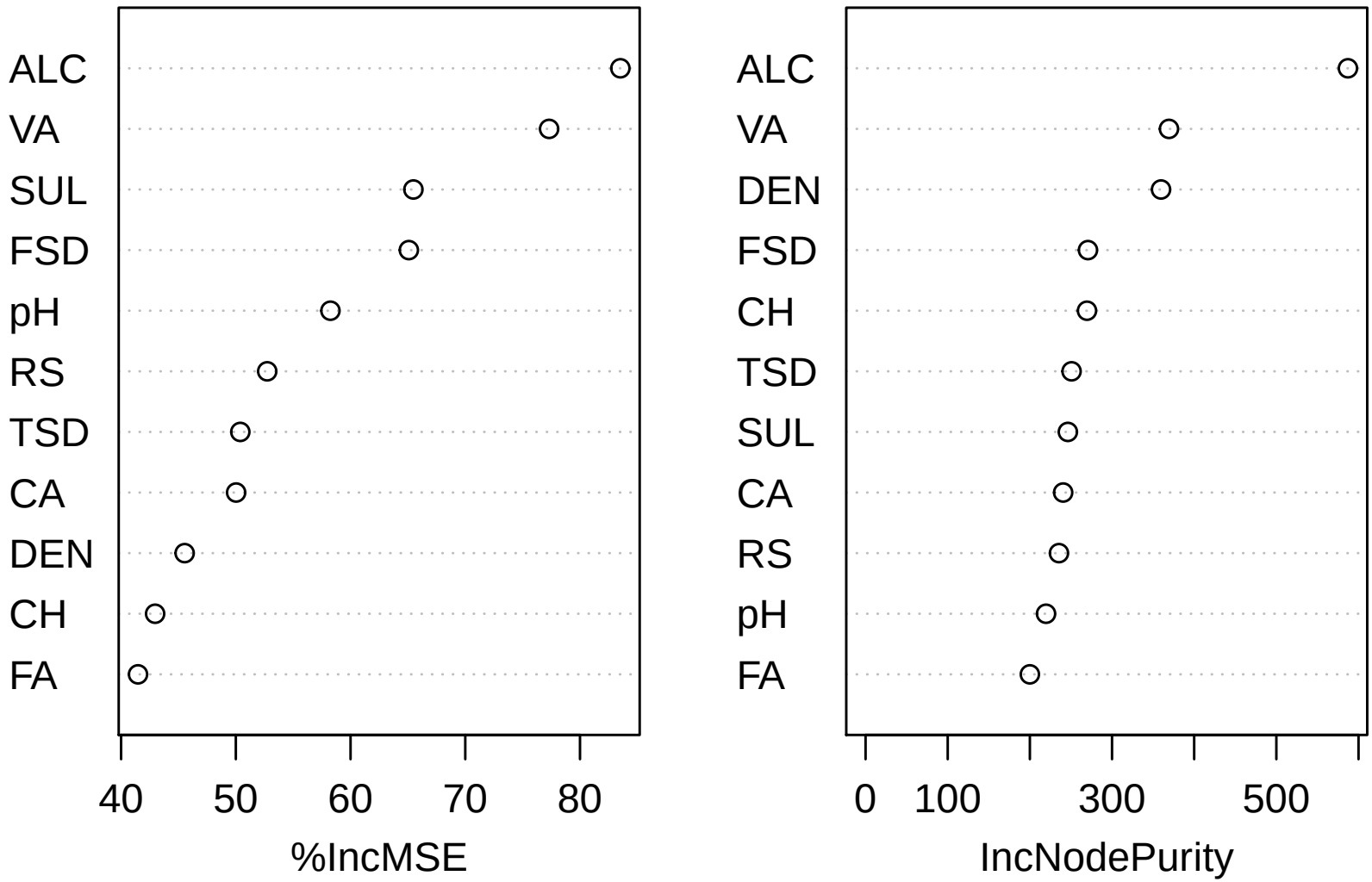
```
library(randomForest)
fitRF1 <- randomForest(
  QLT ~ ., method="anova",
  data=train1.data, importance=TRUE, ntree=500)

PredictionRF1 <- predict(fitRF1, test1.data)

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

```
## [1] 0.7110846
```

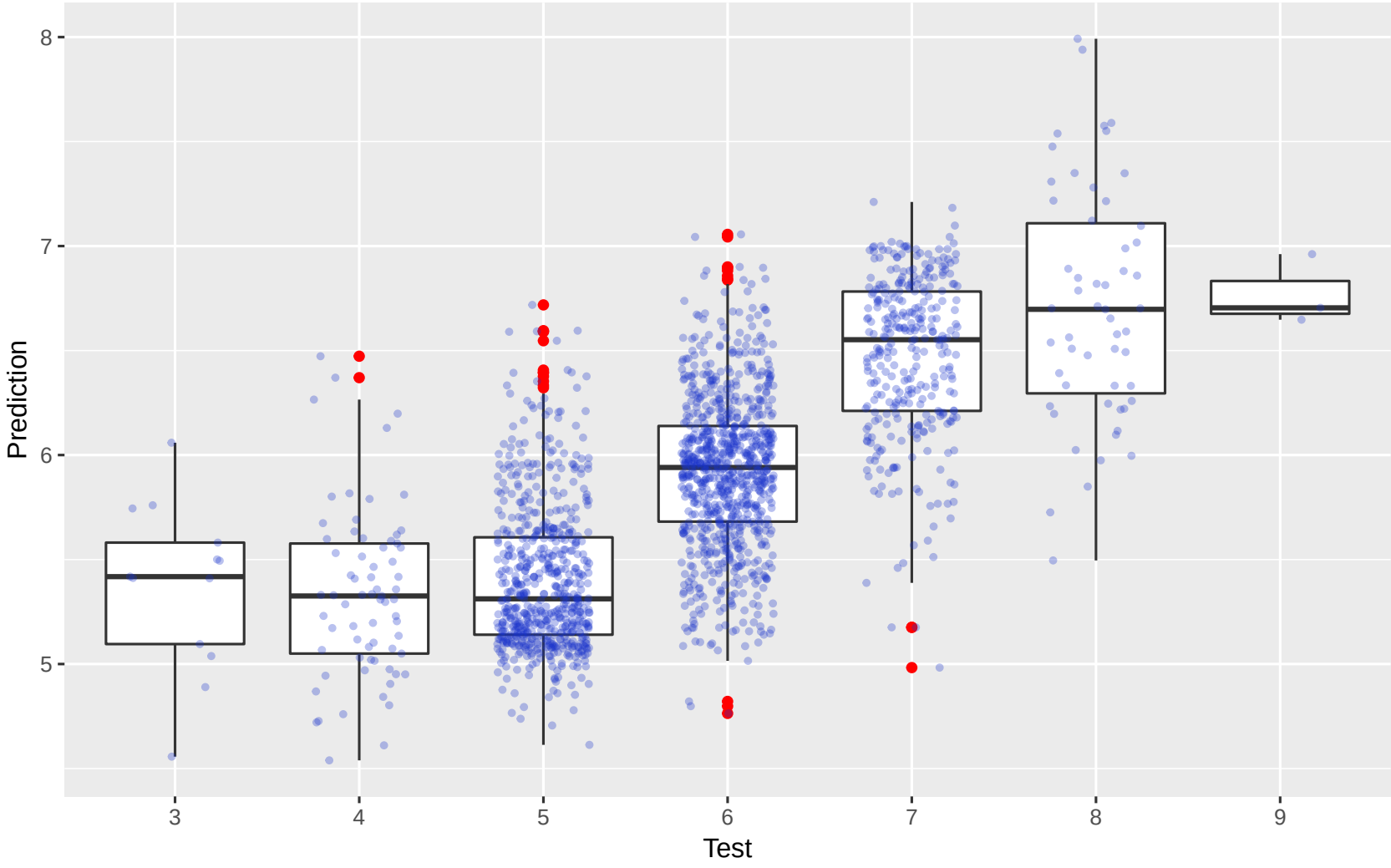
Importance of the dataset attributes for QLT prediction



Random Forest Pledictor Confusion Matrix

	3	4	5	6	7	8	9
5	8	47	424	116	6	1	0
6	5	22	204	687	146	19	0
7	0	0	5	47	173	29	3
8	0	0	0	0	0	6	0

Random Forest Prediction scatter plot



SVM Model and summary

```r
library("e1071")
svm_model <- svm(QLT ~ ., data=train1.data)
summary(svm_model)
```

```
##
## Call:
## svm(formula = QLT ~ ., data = train1.data)
##
##
## Parameters:
##    SVM-Type:  eps-regression
##  SVM-Kernel:  radial
##        cost:  1
##       gamma:  0.09090909
##     epsilon:  0.1
##
##
## Number of Support Vectors:  3880
```

SVM Prediction

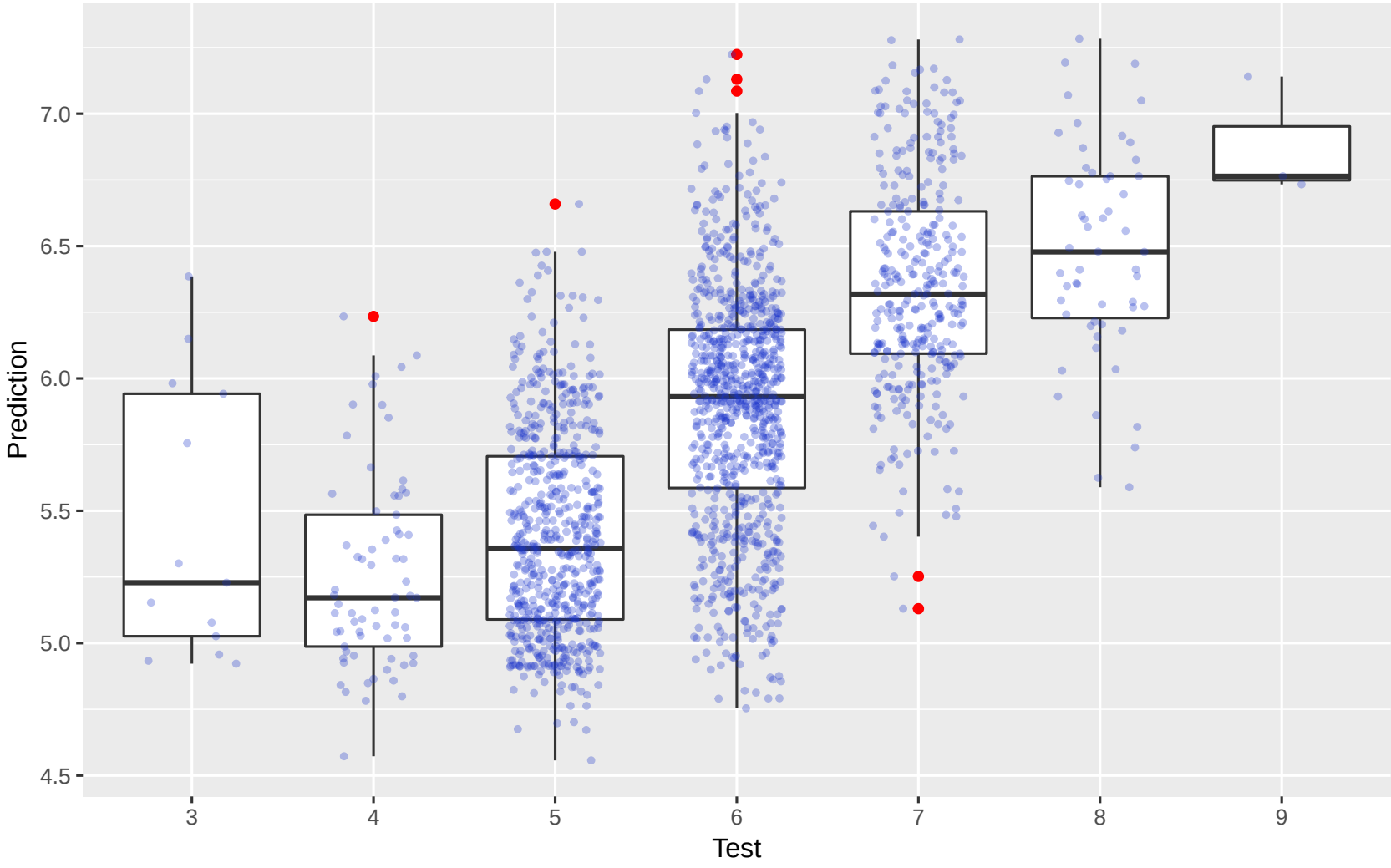
```
predSVM <- predict(svm_model, test1.data)
cor(predSVM,test1.data$QLT)
```

```
## [1] 0.6279396
```

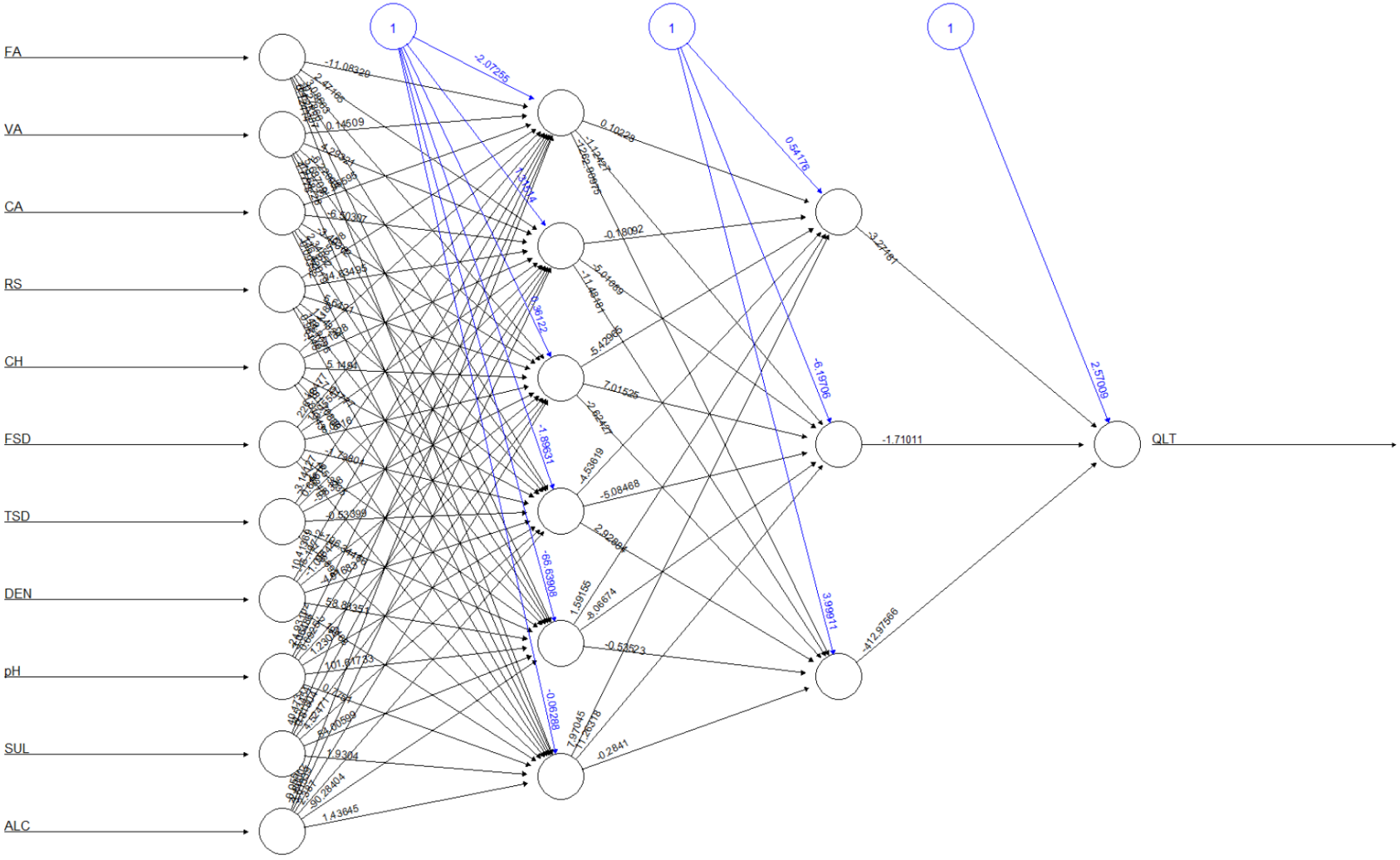
SVM Pledictor Confusion Matrix

	3	4	5	6	7	8	9
5	8	53	394	185	7	0	0
6	5	16	238	599	210	30	0
7	0	0	1	66	108	25	3

SVM Prediction scatter plot



Neural Network Model



Slide with Bullets

- Bullet 1
- Bullet 2
- Bullet 3

Slide with R Output

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

Slide with Plot