«Ανάλυση δεδομένων με σκοπό την διεξαγωγή συμπερασμάτων για την ποιότητα ποικιλίας κόκκινου κρασιού»

Μπατζιάκας Αστέριος

Θεσσαλονίκη, 2014

Περιεχόμενα

1. Περιγραφή του συνόλου δεδομένων & σκοπός της εργασίας

Στην παρούσα εργασία επιλέχθηκε να εξεταστεί ένα σύνολο 1559 παρατηρήσεων που αφορούν διάφορα χαρακτηριστικά του κόκκινου κρασιού και συγκεκριμένα της ποικιλίας *Vinho Verde* που παράγεται και είναι ιδιαίτερα δημοφιλές στην Πορτογαλία. Το όνομα του κυριολεκτικά σημαίνει άγουρο κρασί, που προκύπτει από το ιδιαίτερο χαρακτηριστικό του που είναι το γεγονός ότι οι ζυμώσεις του κρασιού συνεχίζουν και μετά την εμφιάλωση του. Η ποικιλία αυτή του κρασιού περιλαμβάνει κόκκινο, λευκό και ροζέ κρασί. Το Vinho Verde θεωρείται προϊόν προστατευόμενης ονομασίας από την Ευρωπαϊκή Ένωση.

Το αρχικό σύνολο δεδομένων αφορούσε τόσο λευκά όσο και κόκκινα κρασιά και δημιουργήθηκε από μια ομάδα πορτογάλων ερευνητών ανάμεσα στα έτη 2004 και 2007, οι οποίοι προσπάθησαν να εφαρμόσουν μεθόδους machine learning για την δημιουργία μοντέλων με σκοπό την πρόβλεψη της ποιότητας του κρασιού μέσα από την ανάλυση μεταβλητών που ήταν εύκολα μετρήσιμες από απλούς χημικούς ελέγχους. Τα ευρήματα τους είναι δημοσιευμένα στο άρθρο «Modeling wine preferences by data mining from physicochemical properties» που δημοσιεύτηκε στο περιοδικό Decision Support Systems, το 2009 (doi:10.1016/j.dss.2009.05.016). Η χρησιμότητα των μοντέλων αυτών είναι ιδιαίτερα σημαντική καθώς όπως αναφέρουν « *[..] η γεύση είναι η λιγότερο κατανοητή αισθητηριακή λειτουργία*» καθώς και «*επίσης οι σχέσεις ανάμεσα στα φυσιοχημικά και τα αισθητηριακά χαρακτηριστικά είναι περίπλοκες και όχι έως τώρα πλήρως κατανοητές*».

Πλέον όλο το σύνολο δεδομένων είναι δωρεάν διαθέσιμο στον ιστότοπο του πανεπιστημίου της Καλιφόρνια Irvine, το «UCI machine learning repository» στην διεύθυνση: https://archive.ics.uci.edu/ml/datasets/Wine+Quality .

Αντίστοιχα με τους αρχικούς ερευνητές θα προσπαθήσουμε να κατασκευάσουμε δύο μοντέλα σχετικά με τα δεδομένα με σκοπό την πρόβλεψη και την κατηγοριοποίηση των κρασιών σε σχέση με την ποιότητα τους.

2. Περιγραφική στατιστική

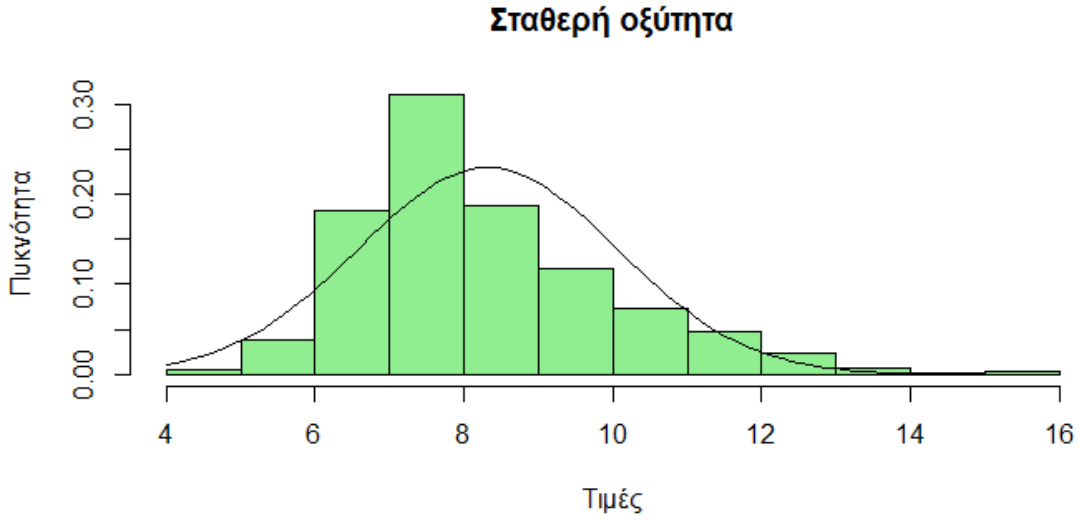
Στο παρόν κομμάτι της εργασίας θα παρουσιαστούν οι δώδεκα μεταβλητές του δείγματος και στην συνέχεια θα προχωρήσουμε στην στατιστική τους ανάλυση, παρουσιάζοντας τα τυπικά στατιστικά μέτρα καθώς και τα σχετικά διαγράμματα.

2.1 Οι μεταβλητές

Σταθερή Οξύτητα

Η μεταβλητή αυτή είναι ποσοτική και μετράει τα γραμμάρια ταρταρικού οξέος ανά κυβικό δεκάμετρο ($^{g}/_{dm^3}$) κρασιού. Στο dataset έχει το όνομα fixed.acidity.

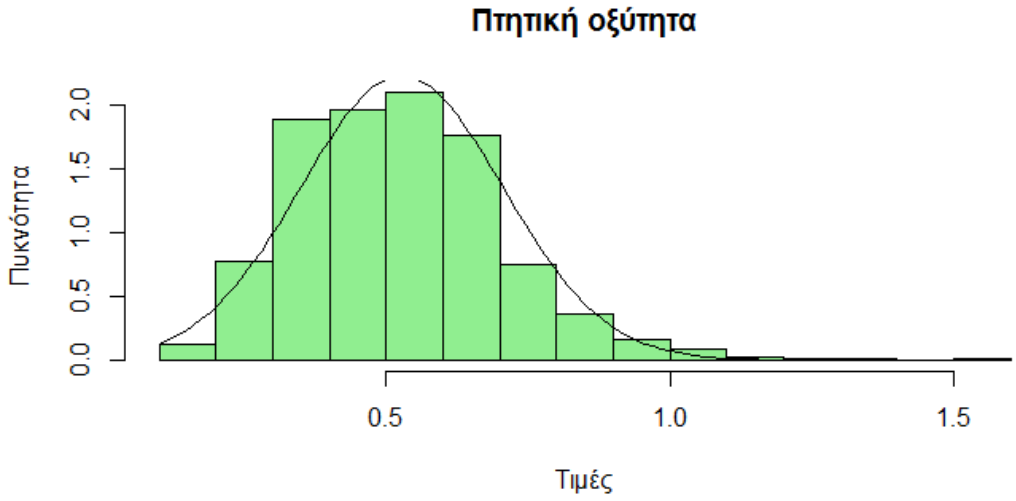
	Σταθερή Οξύτητα
Ελάχιστο	4,6
1ο Τεταρτημόριο	7,1
Διάμεσος	7,9
Μέση Τιμή	8,319637
3ο Τεταρτημόριο	9,2
Μέγιστο	15,9
Διασπορά	3,031416389
Τυπική Απόκλιση	1,741096318



Πτητική Οξύτητα

Η μεταβλητή αυτή είναι ποσοτική και μετράει τα γραμμάρια ακετικού οξέος ανά κυβικό δεκάμετρο ($^{g}/_{dm^3}$) κρασιού. Στο dataset έχει το όνομα volatile.acidity.

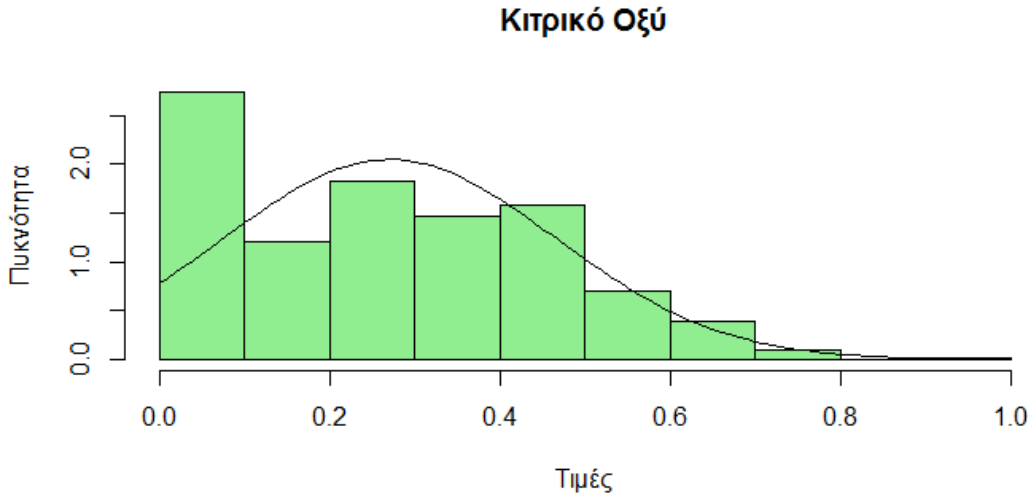
	Πτητική Οξύτητα
Ελάχιστο	0,12
1ο Τεταρτημόριο	0,39
Διάμεσος	0,52
Μέση Τιμή	0,5278205
3ο Τεταρτημόριο	0,64
Μέγιστο	1,58
Διασπορά	0,032062378
Τυπική Απόκλιση	0,179059704



Κιτρικό Οξύ

Η μεταβλητή αυτή είναι ποσοτική και μετράει τα γραμμάρια κιτρικού οξέος ανά κυβικό δεκάμετρο ($^{g}/_{dm^3}$) κρασιού . Στο dataset έχει το όνομα citric.acid.

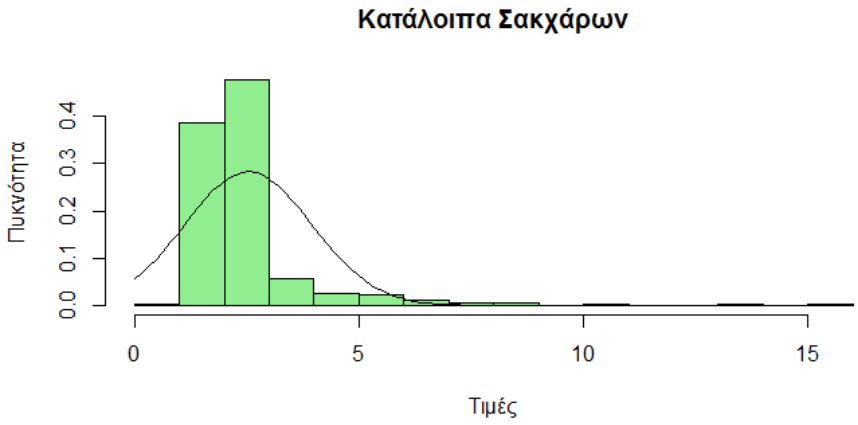
	Κιτρικό Οξύ
Ελάχιστο	0
1ο Τεταρτημόριο	0,09
Διάμεσος	0,26
Μέση Τιμή	0,2709756
3ο Τεταρτημόριο	0,42
	Κιτρικό Οξύ
Μέγιστο	1
Διασπορά	0,037947483
Τυπική Απόκλιση	0,194801137



Κατάλοιπα Σακχάρων

Η μεταβλητή αυτή είναι ποσοτική και μετράει τα γραμμάρια ζακχάρων ανά κυβικό δεκάμετρο ($^{g}/_{dm^3}$) κρασιού . Στο dataset έχει το όνομα residual.sugars.

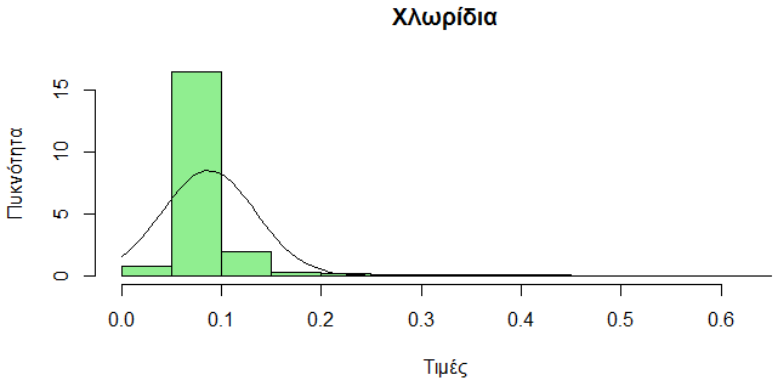
	Κατάλοιπα Σακχάρων
Ελάχιστο	0,9
1ο Τεταρτημόριο	1,9
Διάμεσος	2,2
Μέση Τιμή	2,538806
3ο Τεταρτημόριο	2,6
Μέγιστο	15,5
Διασπορά	1,987897133
Τυπική Απόκλιση	1,40992806



Χλωρίδια

Η μεταβλητή αυτή είναι ποσοτική και μετράει τα γραμμάρια χλωριούχου νάτριου ανά κυβικό δεκάμετρο ($^{g}/_{dm^3}$) κρασιού. Στο dataset έχει το όνομα clorides .

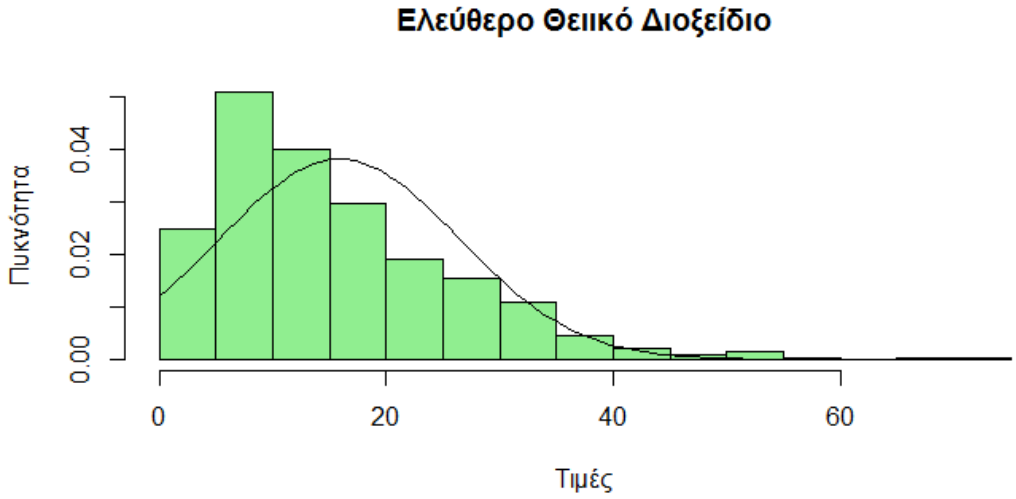
	Χλωρίδια
Ελάχιστο	0,012
1ο Τεταρτημόριο	0,07
Διάμεσος	0,079
Μέση Τιμή	0,08746654
3ο Τεταρτημόριο	0,09
Μέγιστο	0,611
Διασπορά	0,002215143
Τυπική Απόκλιση	0,047065302



Ελεύθερο Θειικό Διοξείδιο

Η μεταβλητή αυτή είναι ποσοτική και μετράει τα μιλιγραμμάρια ελεύθερου Θειικού Διοξειδίου ανά κυβικό δεκάμετρο ($^{mg}/_{dm^3}$) κρασιού. Στο dataset έχει το όνομα free.sulfur.dioxide .

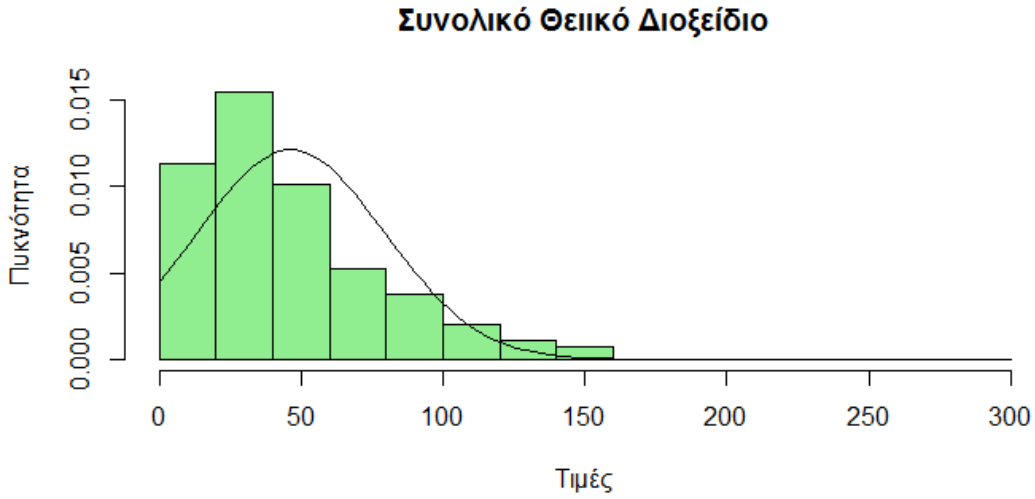
	Ελεύθερο Θειικό Διοξείδιο
Ελάχιστο	1
1ο Τεταρτημόριο	7
Διάμεσος	14
Μέση Τιμή	15,87492
3ο Τεταρτημόριο	21
Μέγιστο	72
Διασπορά	109,415
Τυπική Απόκλιση	10,4602



Συνολικό Θειικό Διοξείδιο

Η μεταβλητή αυτή είναι ποσοτική και μετράει τα μιλιγραμμάρια συνολικού Θειικού Διοξειδίου ανά κυβικό δεκάμετρο ($^{mg}/_{dm^3}$) κρασιού. Στο dataset έχει το όνομα total.sulfur.dioxide .

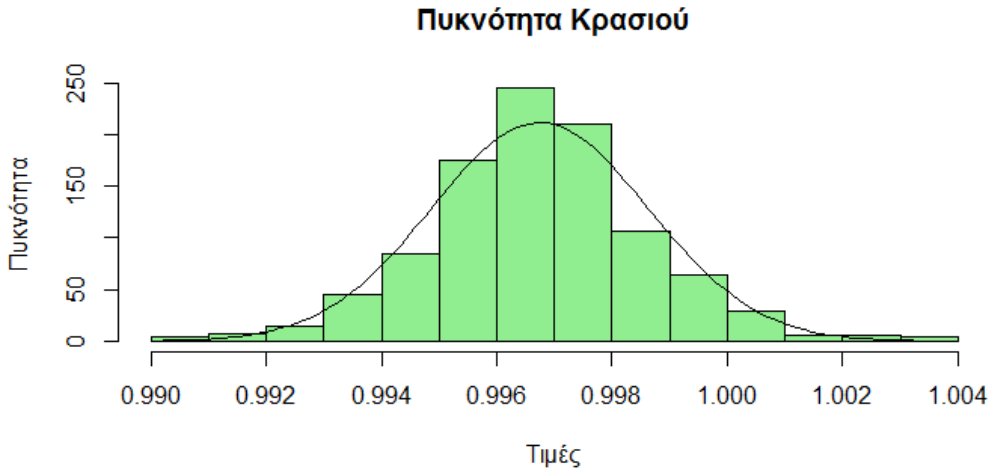
Συνολικό Θειικό Διοξείδιο	
Ελάχιστο	6
1ο Τεταρτημόριο	22
Διάμεσος	38
Μέση Τιμή	46,46779
3ο Τεταρτημόριο	62
Μέγιστο	289
Διασπορά	1082,1
Τυπική Απόκλιση	32,8953



Πυκνότητα

Η μεταβλητή αυτή είναι ποσοτική και μετράει την πυκνότητα του κρασιού ($^{g}/_{cm^3}$) κρασιού. Στο dataset έχει το όνομα density .

	Πυκνότητα
Ελάχιστο	0,99007
1ο Τεταρτημόριο	0,9956
Διάμεσος	0,99675
Μέση Τιμή	0,9967467
3ο Τεταρτημόριο	0,997835
Μέγιστο	1,00369
Διασπορά	3,56E-06
Τυπική Απόκλιση	0,001887334

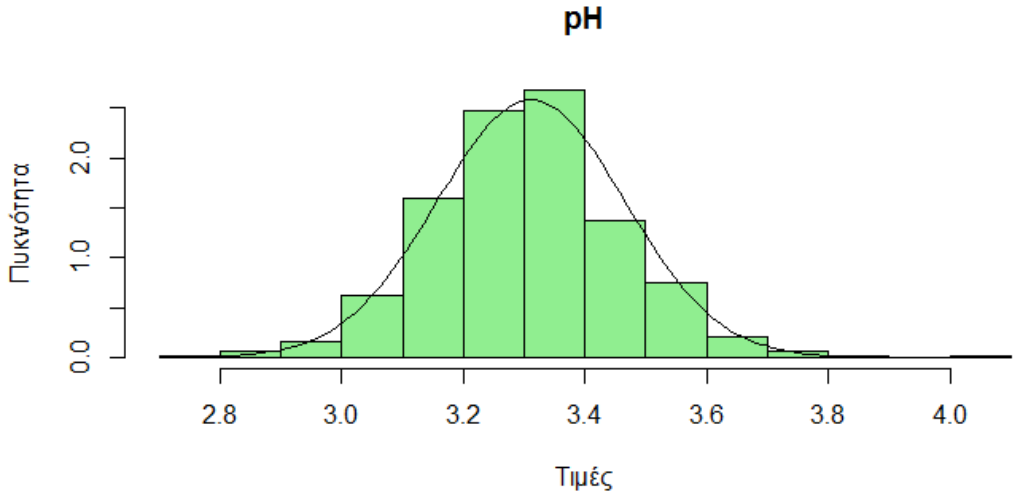


pH

Η μεταβλητή αυτή είναι ποσοτική και αποτελεί αποτελεί μέτρο οξύτητας ή αλκαλικότητας του κρασιού και γενικότερα παίρνει τιμές από 0 έως και 14. Όπως

βλέπουμε στον παρακάτω πίνακα η μέγιστη τιμή είναι 4,01 κάτι που κατατάσσει τα κρασιά που εξετάζουμε στην κατηγορία των όξινων διαλυμάτων. Στο dataset έχει το όνομα pH .

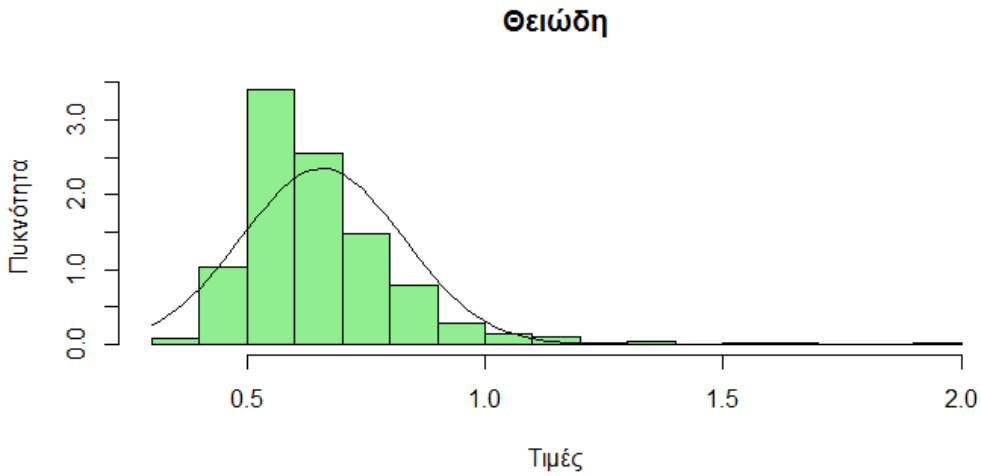
pH	
Ελάχιστο	2,74
1ο Τεταρτημόριο	3,21
Διάμεσος	3,31
Μέση Τιμή	3,311113
3ο Τεταρτημόριο	3,4
Μέγιστο	4,01
Διασπορά	0,023835181
Τυπική Απόκλιση	0,154386465



Θειώδη

Η μεταβλητή αυτή είναι ποσοτική και μετράει τα γραμμάρια θειικού καλίου ανά κυβικό δεκάμετρο ($^{g}/_{dm^3}$) κρασιού. Στο dataset έχει το όνομα sulphates .

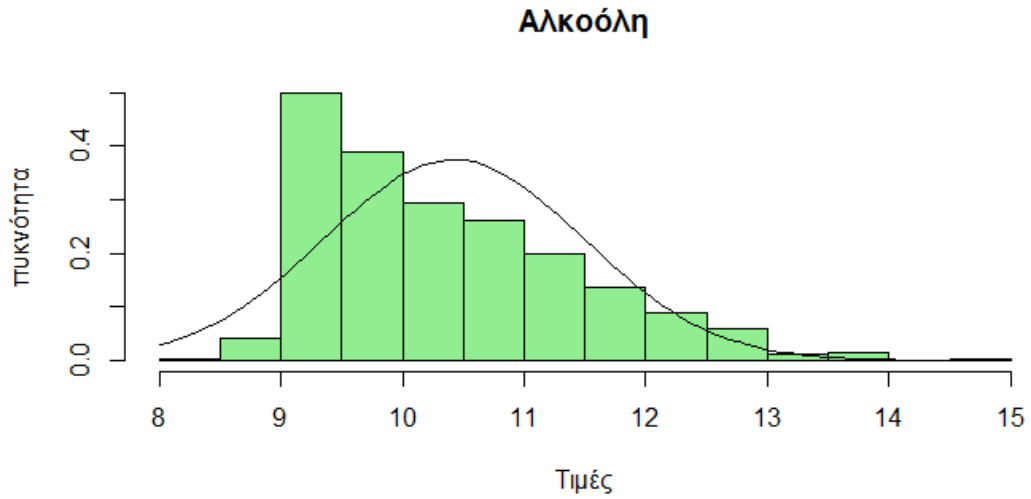
Θειώδη	
Ελάχιστο	0,33
1ο Τεταρτημόριο	0,55
Διάμεσος	0,62
Μέση Τιμή	0,6581488
3ο Τεταρτημόριο	0,73
Μέγιστο	2
Διασπορά	0,028732616
Τυπική Απόκλιση	0,16950698



Αλκοόλη

Η μεταβλητή αυτή είναι ποσοτική και μετράει τους αλκοολικούς βαθμούς του κρασιού δηλαδή την κατ' όγκο περιεκτικότητα της αλκοόλης (vol%) . Στο dataset έχει το όνομα alcohol .

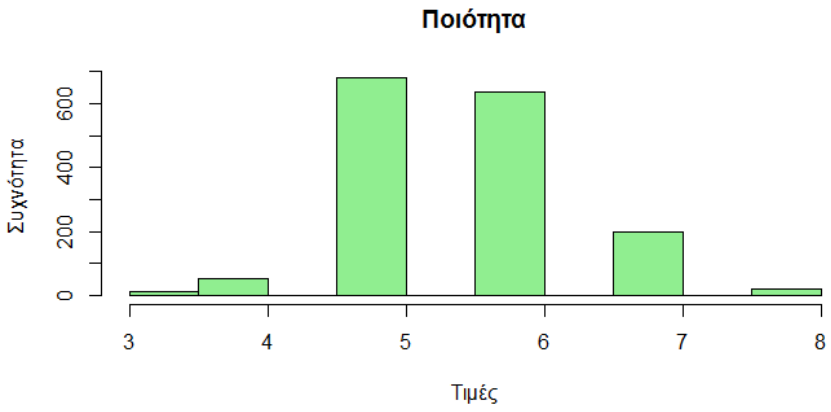
	Αλκοόλη
Ελάχιστο	8,4
1ο Τεταρτημόριο	9,5
Διάμεσος	10,2
Μέση Τιμή	10,42297
3ο Τεταρτημόριο	11,1
Μέγιστο	14,9
Διασπορά	1,13563
Τυπική Απόκλιση	1,06566



Ποιότητα

Η μεταβλητή αυτή είναι η μοναδική ποιοτική του δείγματος. Φανερώνει την ποιότητα του κρασιού σε μια κλίμακα από το 0 (πολύ κακό) έως το 10 (εξαιρετικό). Προέκυψε από τυφλές δοκιμές, όπου τρεις γευσιγνώστες καλέστηκαν να δοκιμάσουν και να αξιολογήσουν το κάθε κρασί και είναι ο μέσος όρος στρογγυλοποιημένος προς τον κάτω ακέραιο των τριών βαθμολογιών. Παρά το γεγονός ότι η μεταβλητή είναι ποιοτική, λόγω της φύσης της κλίμακας υπάρχει νόημα να την εξετάσουμε και ως ποσοτική. Παρατηρούμε πως από το δείγμα απουσιάζουν κρασιά ποιότητας 0,1,2,9 και 10, ενώ πλειοψηφία αποτελούν αυτά με ποιότητα 5 και 6 (≈82,6%).

Ποιότητα	
Ελάχιστο	3
1ο Τεταρτημόριο	5
Διάμεσος	6
Μέση Τιμή	5,636023
3ο Τεταρτημόριο	6
Μέγιστο	8
Διασπορά	0,6521684
Τυπική Απόκλιση	0,80756944



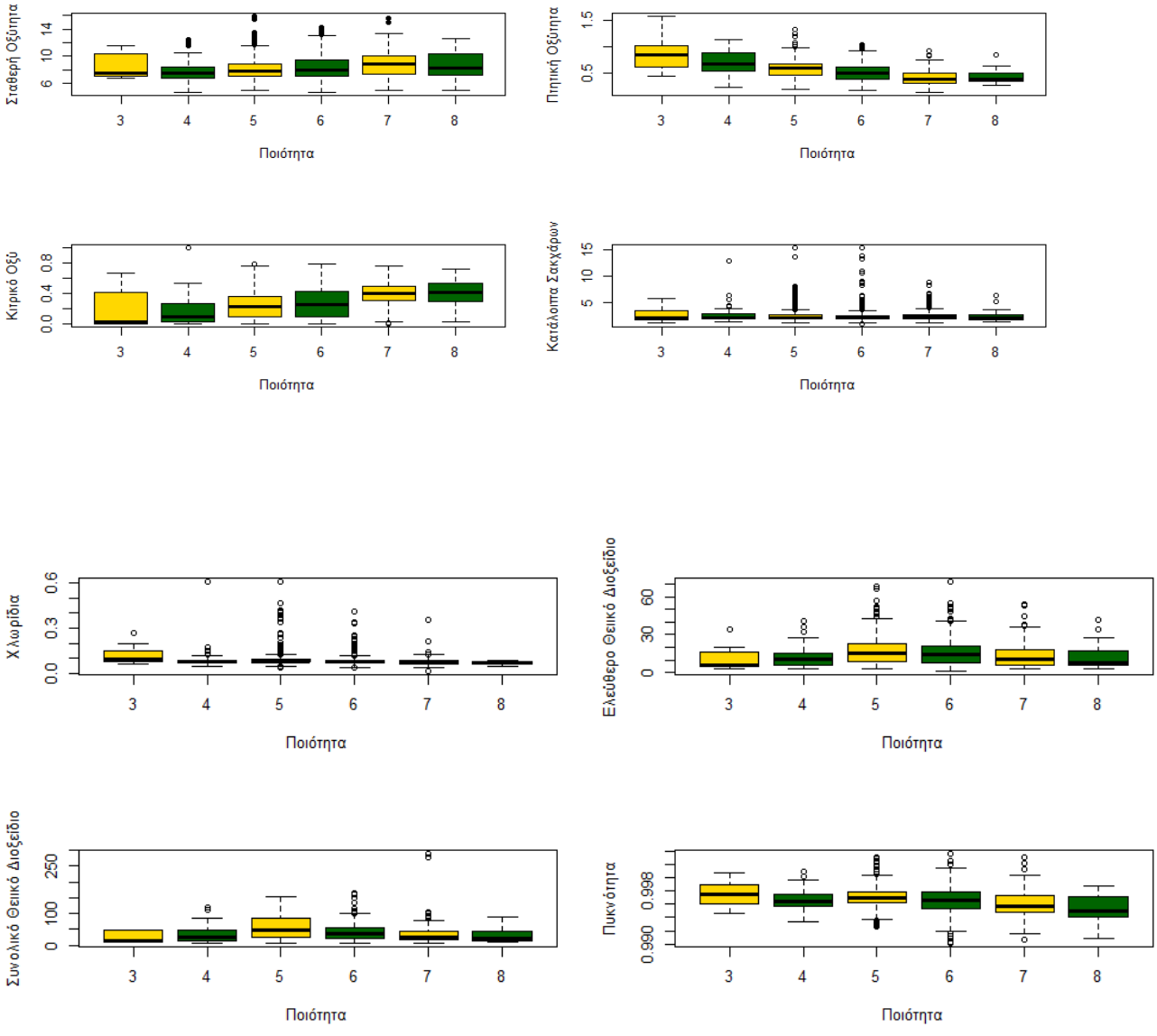
	Συχνότητα	Ποσοστό
Ποιότητα 3	10	0,625%
Ποιότητα 4	53	3,315%
Ποιότητα 5	681	42,589%
Ποιότητα 6	638	39,900%
Ποιότητα 7	199	12,445%
Ποιότητα 8	18	1,126%
Άθροισμα	1599	100,000%

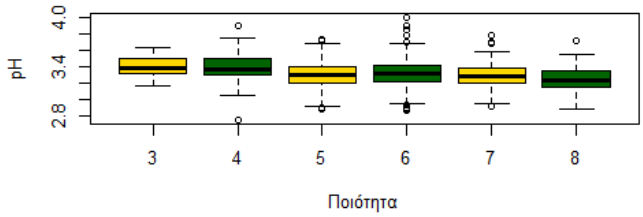
2.2 Θηκογράμματα

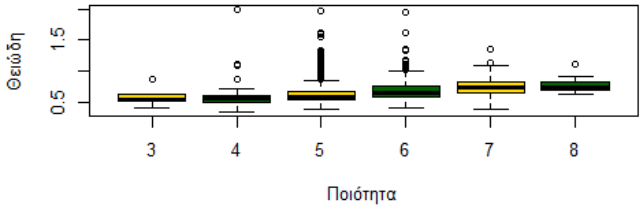
Με την κατασκευή των θηκογραμμάτων μπορούμε να βγάλουμε κάποια πρώτα συμπεράσματα σχετικά με το πώς επηρεάζεται η ποιότητα του κρασιού από τις διάφορες μεταβλητές:

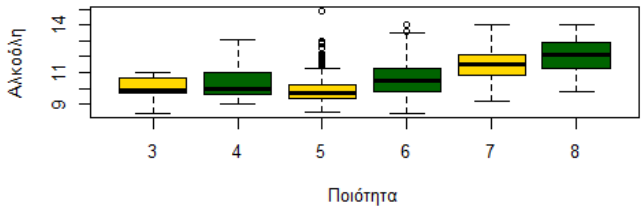
- Τα κρασιά με καλύτερη ποιότητα περιέχουν περισσότερο κιτρικό οξύ και θειώδη καθώς και έχουν μεγαλύτερη περιεκτικότητα σε αλκοόλη.
- Φαίνεται να υπάρχει αντιστρόφως ανάλογη σχέση ανάμεσα στην πηκτική οξύτητα και την ποιότητα
- Τα κρασιά με ποιότητα μεγαλύτερη ή ίση με 5 είναι λιγότερα όξινα σε σχέση με αυτά που είναι ποιότητα 3 και 4

Τέλος παρατηρούμε πως υπάρχει μεγάλο πλήθος ακραίων τιμών (outliers) σε όλες τις μεταβλητές εκτός από το κιτρικό οξύ.









2.3 Συσχέτιση των μεταβλητών

	Σταθερή Οξύτητα	Πτητική Οξύτητα	Κιτρικό Οξύ	Κατάλοιπα Σακχάρων	Χλωρίδια	Ελεύθερο Θειικό Διοξείδιο	Συνολικό Θειικό Διοξείδιο	Πυκνότητα	pH	Θειώδη	Αλκοόλη
Σταθερή Οξύτητα	1	-0,256	0,672	0,115	0,094	-0,154	-0,113	0,261	-0,683	0,183	-0,057
Πτητική Οξύτητα	-0,256	1	-0,552	0,002	0,061	-0,011	0,076	0,005	0,235	-0,261	0,118
Κιτρικό Οξύ	0,672	-0,552	1	0,144	0,204	-0,061	0,036	0,088	-0,542	0,313	-0,093
Κατάλοιπα Σακχάρων	0,115	0,002	0,144	1	0,056	0,187	0,203	0,299	-0,086	0,006	-0,048
Χλωρίδια	0,094	0,061	0,204	0,056	1	0,006	0,047	0,074	-0,265	0,371	0,141
Ελεύθερο Θειικό Διοξείδιο	-0,154	-0,011	-0,061	0,187	0,006	1	0,668	0,094	0,07	0,052	0,052
Συνολικό Θειικό Διοξείδιο	-0,113	0,076	0,036	0,203	0,047	0,668	1	0,039	-0,066	0,043	0,229
Πυκνότητα	0,261	0,005	0,088	0,299	0,074	0,094	0,039	1	-0,125	0,048	0,033
pH	-0,683	0,235	-0,542	-0,086	-0,265	0,07	-0,066	-0,125	1	-0,197	-0,079
Θειώδη	0,183	-0,261	0,313	0,006	0,371	0,052	0,043	0,048	-0,197	1	-0,065
Αλκοόλη	-0,057	0,118	-0,093	-0,048	0,141	0,052	0,229	0,033	-0,079	-0,065	1

Ο παραπάνω πίνακας περιέχει τους συντελεστές συσχέτισης για όλες τις μεταβλητές εκτός από την ποιότητα. Παρατηρούμε ότι:

- Ισχυρή θετική συσχέτιση έχουν οι μεταβλητές: Κιτρικό οξύ-Σταθερή οξύτητα και Ελεύθερο Θειικό Διοξείδιο - Συνολικό Θειικό Διοξείδιο

- Ισχυρή αρνητική συσχέτιση έχουν οι μεταβλητές: Σταθερή οξύτητα – pH και Πτητική οξύτητα – Κιτρικό οξύ.

3. Παλινδρόμηση και SVM

3.1 Προετοιμασία των δεδομένων

Βλέποντας το πλήθος των ακραίων τιμών στο δείγμα, έγινε η επιλογή στην συνέχεια της εργασίας να δημιουργηθεί ακόμα ένα dataset που δεν θα τις περιλαμβάνει και να εφαρμοστεί και στα δύο η ίδια μοντελοποίηση ώστε να έχουμε συγκρίσιμα αποτελέσματα.

Η αποκοπή από τιμών από το δείγμα έγινε σύμφωνα με τον τύπο x>Q3+1.5IQR, δηλαδή όσες μεταβλητές είχαν τιμή μεγαλύτερη από το άθροισμα της τιμής του τρίτου τεταρτημορίου κάθε μεταβλητής και του εύρους των τεταρτημορίων πολλαπλασιασμένου κατά 1,5. Το νέο σύνολο δεδομένων είχε 1212 παρατηρήσεις, δηλαδή 387 λιγότερες από το αρχικό. Στην συνέχεια δημιουργήθηκαν 4 υποσύνολα δεδομένων με σκοπό την εκπαίδευση και την δοκιμασία των μοντέλων σε αντιστοιχία 0.7 προς 0.3 αντίστοιχα.

3.2 Πολλαπλή Παλινδρόμηση

Θα ξεκινήσουμε εργαζόμενοι στο μοντέλο με τα outliers. Από τα συμπεράσματα που αναπτύχθηκαν παραπάνω με βάση τα θηκογράμματα, πάρθηκε η επιλογή ο τύπος παλινδρόμησης που θα εξεταστεί να περιέχει και την σχέση της αλκοόλης με τις υπόλοιπες μεταβλητές. Αντίστοιχα θα μπορούσε να χρησιμοποιηθεί και το κιτρικό οξύ. Πράγματι, μπορεί κάποιος να διαπιστώσει πως ένα τέτοιο μοντέλο δίνει καλύτερα αποτελέσματα σε σχέση με ένα απλούστερο μοντέλο πολλαπλής παλινδρόμησης.Στην συνέχεια, με την βοήθεια της εντολής step(), εντοπίστηκε το μοντέλο με το χαμηλότερο βαθμό στο κριτήριο Akaike, το οποίο είναι:

```
Step:  AIC=-1399.87
quality ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide +
    pH + chlorides + free.sulfur.dioxide + citric.acid + alcohol:sulphates
```

```
                                Df Sum of Sq    RSS     AIC
<none>                                      657.97 -1399.9
- citric.acid                    1     0.840 658.81 -1399.8
+ alcohol:chlorides              1     0.374 657.59 -1398.8
+ alcohol:free.sulfur.dioxide    1     0.272 657.70 -1398.5
+ alcohol:pH                     1     0.269 657.70 -1398.5
+ residual.sugar                 1     0.261 657.71 -1398.5
+ alcohol:volatile.acidity       1     0.081 657.89 -1398.1
+ alcohol:citric.acid            1     0.043 657.93 -1398.0
+ alcohol:total.sulfur.dioxide   1     0.042 657.93 -1398.0
+ fixed.acidity                  1     0.015 657.95 -1397.9
- free.sulfur.dioxide            1     1.642 659.61 -1397.9
- chlorides                      1     4.352 662.32 -1391.3
- pH                             1     7.330 665.30 -1384.2
- alcohol:sulphates              1     9.103 667.07 -1379.9
- total.sulfur.dioxide          1     9.464 667.43 -1379.0
- volatile.acidity              1    36.396 694.37 -1315.8
```

Το μοντέλο αυτό έχει σημαντικά μικρότερο AIC από το αντίστοιχο που περιέχει απλά άθροιση όλων των μεταβλητών (AIC=-682.5).

Από την σύνοψη του μοντέλου μπορούμε να βγάλουμε κάποια συμπεράσματα:

```
Residuals:
     Min       1Q   Median       3Q      Max
-2.41623 -0.38582 -0.07449  0.43867  2.11561

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  5.3885438  0.9077353   5.936  3.9e-09 ***
volatile.acidity            -1.1725566  0.1177384  -9.959  < 2e-16 ***
alcohol                      0.0381704  0.0890183   0.429  0.66816
sulphates                   -2.1713631  1.3010531  -1.669  0.09541 .
chlorides                   -1.3444118  0.4718892  -2.849  0.00447 **
total.sulfur.dioxide         0.0051506  0.0078648   0.655  0.51267
residual.sugar               0.0182302  0.0141159   1.291  0.19681
free.sulfur.dioxide         -0.0354475  0.0261596  -1.355  0.17568
alcohol:sulphates            0.2976635  0.1265659   2.352  0.01885 *
alcohol:free.sulfur.dioxide  0.0037562  0.0025229   1.489  0.13681
alcohol:total.sulfur.dioxide -0.0007991  0.0007648  -1.045  0.29631
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6483 on 1109 degrees of freedom
Multiple R-squared:  0.3509,   Adjusted R-squared:  0.345
F-statistic: 59.94 on 10 and 1109 DF,  p-value: < 2.2e-16
```

Το p –value είναι μικρότερο του 0.05 αρά επιβεβαιώνεται η υπόθεση πως τουλάχιστον μια από τις μεταβλητές του μοντέλου είναι σημαντική για την πρόβλεψη της ποιότητας ενώ από το MRS βλέπω πως εξηγείται το 35,09 της απόκλισης. Ωστόσο στην τελευταία στήλη βλέπω πως οι περισσότερες από τις μεταβλητές δεν έχουν στατιστική σημασία για το μοντέλο (τιμή <0.05).

Ελέγχοντας την επιτυχία του μοντέλου διαπιστώνω πως έχει ποσοστό επιτυχίας 69,31% .

Συνεχίζοντας αρχίσαμε να αφαιρούμε κάθε φορά την μεταβλητή με το μεγαλύτερο p. Με την σειρά εξαλείφθηκαν από το μοντέλο οι μεταβλητές: Αλκοόλη, Συνολικό Θειικό Διοξείδιο, κατάλοιπα σακχάρων, Ελεύθερο Θειικό Διοξείδιο και Αλκοόλη* Ελεύθερο Θειικό Διοξείδιο. Το τελικό μοντέλο έχει συνοπτικό πίνακα:

```
Residuals:
    Min      1Q  Median      3Q     Max
-2.3791 -0.3901 -0.0729  0.4462  2.1118
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  5.823e+00  1.139e-01  51.101  < 2e-16 ***
volatile.acidity            -1.177e+00  1.173e-01 -10.036  < 2e-16 ***
sulphates                   -3.251e+00  3.338e-01  -9.740  < 2e-16 ***
chlorides                   -1.259e+00  4.550e-01  -2.768 0.005739 **
sulphates:alcohol            4.042e-01  2.802e-02  14.427  < 2e-16 ***
alcohol:total.sulfur.dioxide -2.142e-04  5.912e-05  -3.623 0.000304 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6485 on 1114 degrees of freedom
Multiple R-squared:  0.3474,   Adjusted R-squared:  0.3445
F-statistic: 118.6 on 5 and 1114 DF,  p-value: < 2.2e-16
```

Πλέον όλες οι μεταβλητές θεωρούνται σημαντικές. Ο τύπος του μοντέλου είναι ο εξής:

$$\text{Ποιότητα} = 5{,}8227 - 1.1772 \times \text{Πτητική Οξύτητα} - 3.2513 \times \text{Θειώδη} - 1.2592 \times \text{Χλωρίδια} + 0.4041 \times \text{Θειώδη} \times \text{Αλκοόλη} - 0.0002 \times \text{Αλκοόλη} \times \text{Συνολικό θειικό Διοξείδιο}$$

Το ποσοστό επιτυχίας του μοντέλου είναι 69,93% .

Για το σύνολο των δεδομένων χωρίς ακραίες τιμές, ακολουθήσαμε ακριβώς την ίδια διαδικασία:

```
quality ~ sulphates + volatile.acidity + alcohol +
    chlorides + total.sulfur.dioxide + pH + residual.sugar +
    free.sulfur.dioxide + fixed.acidity, data = train.no)

AIC=-1240.97


Residuals:
     Min       1Q   Median       3Q      Max
-2.05940 -0.34894 -0.06506  0.41628  1.91763

Coefficients:
                      Estimate Std. Error t value             Pr(>|t|)
(Intercept)           4.022787   0.855574   4.702           0.00000301 ***
sulphates             1.664224   0.194727   8.546 < 0.0000000000000002 ***
volatile.acidity     -0.611659   0.140605  -4.350           0.00001528 ***
alcohol               0.309291   0.024020  12.876 < 0.0000000000000002 ***
chlorides            -1.223322   1.573383  -0.778              0.43708
total.sulfur.dioxide -0.003367   0.001073  -3.139              0.00175 **
pH                   -0.634266   0.216579  -2.929              0.00350 **
residual.sugar        0.027811   0.052726   0.527              0.59801
free.sulfur.dioxide   0.006042   0.003112   1.942              0.05249 .
fixed.acidity        -0.016390   0.021087  -0.777              0.43722
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6008 on 839 degrees of freedom
Multiple R-squared:  0.3913,   Adjusted R-squared:  0.3848
F-statistic: 59.93 on 9 and 839 DF,  p-value: < 0.00000000000000022
```

Παρατηρούμε πως σε σχέση με τα προηγούμενα μοντέλα έχει μειωθεί το p-value του F-statistic καθώς και το MRS και έχουμε λιγότερες μη στατιστικά σημαντικές μεταβλητές. Ακόμη έχουμε σημαντική αύξηση του ποσοστού πρόβλεψης, στο 72,17%. Τέλος σε αντίθεση με το παραπάνω μοντέλο δεν υπάρχουν σχέσεις μεταξύ των μεταβλητών.

Αφαιρώ με την ίδια διαδικασία που περιγράφηκε παραπάνω τις μεταβλητές: Κατάλοιπα Σακχάρων, Σταθερή οξύτητα, χλωρίδια για να καταλήξω στο τελικό μου μοντέλο, με πίνακα σύνοψης:

```
Residuals:
     Min       1Q   Median       3Q      Max
-2.05311 -0.35254 -0.06342  0.42094  1.92332

Coefficients:
                      Estimate Std. Error t value             Pr(>|t|)
(Intercept)           3.337227   0.544463   6.129           0.00000000135 ***
sulphates             1.623585   0.188843   8.598 < 0.0000000000000002 ***
```

```
volatile.acidity     -0.603358   0.138192  -4.366        0.00001422409 ***
alcohol               0.317144   0.022712  13.964 < 0.0000000000000002 ***
total.sulfur.dioxide -0.003303   0.001040  -3.176              0.00155 **
pH                   -0.499080   0.161228  -3.095              0.00203 **
free.sulfur.dioxide   0.006331   0.003094   2.046              0.04106 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6003 on 842 degrees of freedom
Multiple R-squared:  0.3903,   Adjusted R-squared:  0.386
F-statistic: 89.84 on 6 and 842 DF,  p-value: < 0.00000000000000022
```

Ο τύπος του μοντέλου είναι:

$$Ποιότητα = 3.337227 + 1.624 \times \text{Θειώδη} - 0.603358 \times \text{Πτητική Οξύτητα} + 0.317144 \times \text{Αλκοόλη} - 0.0033 \times \text{Ελεύθερο Θειικό Διοξείδιο} - 0.499 \times \text{pH} + 0.0063 \times \text{Ελεύθερο Θειικό Διοξείδιο}$$

Με ποσοστό επιτυχίας 71,9%, μικρότερο του αρχικού.

3.3 Support Vector Machines

Για το δύο σύνολα δεδομένων αρχικά δοκιμάστηκε η κατασκευή SVM με διαφορετικούς πυρήνες με σκοπό να δούμε ποιος έχει την μεγαλύτερη ευστοχία (accuracy):

Πυρήνας	Dataset με Outliers	Dataset χωρίς Outiers
Linear	73.14%	73,75%
Polynomial	71.61%	73.30%
Radial	74.44%	75.80%

Στην συνέχεια με 10-cross validation έγινε η προσπάθεια να εντοπιστούν οι καλύτερες τιμές για το cost και το gamma, στα πλαίσια της διαθέσιμης υπολογιστικής δύναμης. Αυτή η διαδικασία ήταν η πιο απαιτητική σε ολόκληρη την εργασία καταλήγοντας σε runtime 4-5 λεπτών. Εν τέλει καταλήξαμε στα εξής μοντέλα ταξινόμησης:

Dataset	Πυρήνας	Cost	Gamma
Με outliers	Radial	1	0.5
Χωρίς outliers	Radial	1	0.5

Confusion Matrixes

Για το set με τα outliers:

Ποιότητα	Μέτρια	Καλή
Μέτρια	217	43
Καλή	57	162

Ποσοστό επιτυχημένων προβλέψεων:

Για το set χωρίς τα outliers: 79,12%

Ποιότητα	Μέτρια	Καλή
Μέτρια	174	49
Καλή	32	108

Ποσοστό επιτυχημένων προβλέψεων: 77,68%

Τέλος με την βοήθεια της εντολής confusionMatrix() από την βιβλιοθήκη caret παίρνουμε τα εξής αποτελέσματα:

Σετ με outliers

```
Accuracy : 0.762

95% CI : (0.7213, 0.7995)
    No Information Rate : 0.572
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.5192
 Mcnemar's Test P-Value : 0.1113

            Sensitivity : 0.7591
            Specificity : 0.7659
         Pos Pred Value : 0.8125
         Neg Pred Value : 0.7040
             Prevalence : 0.5720
         Detection Rate : 0.4342
   Detection Prevalence : 0.5344
      Balanced Accuracy : 0.7625

       'Positive' Class : Καλή
```

Σετ χωρίς Outliers

```
 Accuracy : 0.7769

95% CI : (0.7305, 0.8187)
    No Information Rate : 0.5675
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.5395
 Mcnemar's Test P-Value : 0.07544

            Sensitivity : 0.8447
            Specificity : 0.6879
         Pos Pred Value : 0.7803
         Neg Pred Value : 0.7714
             Prevalence : 0.5675
         Detection Rate : 0.4793
   Detection Prevalence : 0.6143
      Balanced Accuracy : 0.7663

       'Positive' Class : Καλή
```