

Metodi Informatici per la Gestione Aziendale - Relazione Finale Progetto

Giuseppe Facchi

A.A. 2020-2021

Indice

1	Panoramica	3
2	Data Acquisiton	4
2.1	Descrizione del dataset	4
2.2	Descrizione dei valori delle variabili	5
2.3	Descrizione delle distribuzioni delle variabili	7
2.3.1	Qualità	7
2.3.2	Acidi non volatili	8
2.3.3	Acidi volatili	9
2.3.4	Acido citrico	10
2.3.5	Zuccheri residui	11
2.3.6	Cloruri	12
2.3.7	Anidride Solforosa libera	13
2.3.8	Anidride Solforosa totale	14
2.3.9	Densità	15
2.3.10	pH	16
2.3.11	Solfati	17
2.3.12	Percentuale di Alcol	18
2.4	Correlazione tra variabili	19
3	Regressione e Classificazione	20
3.1	Organizzazione di training e test set	20
3.2	Regressione	20
3.2.1	Regressione Lineare Multipla	20
3.2.2	Albero di regressione semplice	22
3.2.3	M5P	22
3.3	Classificazione	24
3.3.1	Introduzione	24
3.3.2	Albero di classificazione semplice	25
3.3.3	Random Forest	27
4	Conclusioni	32

1 Panoramica

Lo scopo principale dell'applicazione vertono sulla predizione e sulla classificazione della qualità di varianti del vino bianco Portoghese *vinho verde* in base alle loro proprietà chimiche. Dopo l'ottenimento di accettabili modelli di **regressione** e di **classificazione**, sarà possibile effettuare analisi di mercato per poter produrre varianti di vino sempre più perfezionate e progettare marketing strategies basate sul target di utenza da soddisfare.

Vinho Verde è una denominazione che annovera tipologie non solo di vino bianco, ma anche di vino rosso e rosato. Viene prodotto in tutto il Portogallo da monovitigni o in uvaggio, è leggermente frizzante e presenta aromi di miele, di frutta e di fiori di campo.



Figura 1: Varianti di Vinho Verde portoghese

2 Data Acquisiton

2.1 Descrizione del dataset

Il dataset utilizzato è reperibile a questo link. Tra le due collezioni è stata esaminata quella relativa alle varianti di vinho verde bianco. Di seguito le sue principali caratteristiche:

- N° totale di osservazioni: **4898**
- N° di variabili: **12**

Ogni variante di vinho verde è descritta tramite valori di

- **Acidi Non Volatili:** Gli acidi non volatili sono composti chimici che non possono essere rapidamente vaporizzati. Ciò è principalmente dovuto al fatto che la pressione di vapore dell'acido a temperatura ambiente normale non è sufficientemente elevata da vaporizzare facilmente
- **Acidi Volatili:** Gli acidi volatili sono composti chimici che subiscono rapidamente la vaporizzazione. Questa rapida vaporizzazione è il risultato di una pressione di vapore elevata a temperatura ambiente normale. Pertanto, gli acidi volatili possono subire vaporizzazione senza riscaldamento o qualsiasi altra forza esterna
- **Acido Citrico:** Viene utilizzato nel vino per prevenire le ossidazioni e si presenta, come additivo o acidificante, sotto forma di polverina bianca e cristallina pura
- **Zuccheri Residui:** Determinano la maggiore o minore dolcezza del vino
- **Cloruri:** Determinano la sensazione di sapidità
- **Anidride Solforosa Libera:** Forme liberabili in seguito ad acidificazione
- **Anidride Solforosa Totale:** Usata in giuste quantità migliora la qualità e prolunga la durata del vino
- **Densità**

- **PH:** Indica l'acidità/basicità) del vino
- **Solfiti:** Funzione principale di conservanti, per contrastare l'ossidazione dei cibi e prevenire lo sviluppo microbico indesiderato
- **% Alcol:** Numero di parti in volume di alcol etilico, alla temperatura di 20 °C, contenuta in 100 parti in volume del prodotto considerato alla stessa temperatura
- **Qualità:** Valutazione soggettiva del vino da 1 a 10

2.2 Descrizione dei valori delle variabili

Come evidenziato dalla seguente infografica non sono presenti missing values nel dataset in analisi. Sarà quindi possibile proseguire l'analisi senza effettuare operazioni di modifica dell'integrità del dataset.

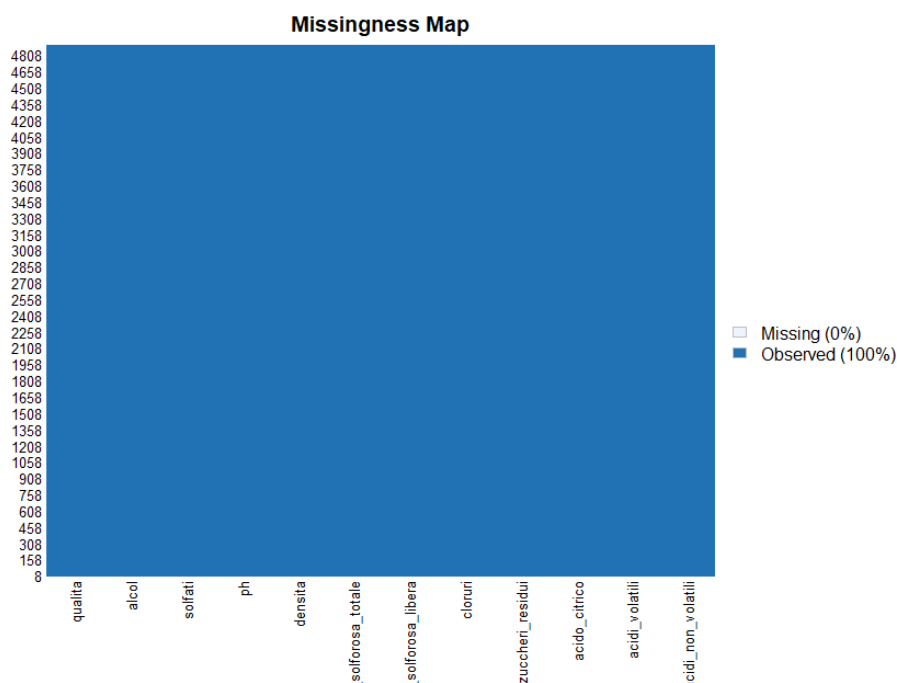


Figura 2: Missmap del dataset in analisi

Una panoramica generale dei valori del dataset è illustrata di seguito. Tutte le variabili sono rappresentate da dati numerici:

```

$ fixed.acidity      : num  6.7 5.7 5.9 5.3 6.4 7 7.9 6.6 7 6.5 ...
$ volatile.acidity  : num  0.62 0.22 0.19 0.47 0.29 0.14 0.12 0.38 0.16 0.37 ...
$ citric.acid       : num  0.24 0.2 0.26 0.1 0.21 0.41 0.49 0.28 0.3 0.33 ...
$ residual.sugar    : num  1.1 16 7.4 1.3 9.65 0.9 5.2 2.8 2.6 3.9 ...
$ chlorides         : num  0.039 0.044 0.034 0.036 0.041 0.037 0.049 0.043 0.043 0.027 ...
$ free.sulfur.dioxide : num  6 41 33 11 36 22 33 17 34 40 ...
$ total.sulfur.dioxide : num  62 113 123 74 119 95 152 67 90 130 ...
$ density           : num  0.993 0.999 0.995 0.991 0.993 ...
$ pH               : num  3.41 3.22 3.49 3.48 2.99 3.25 3.18 3.21 2.88 3.28 ...
$ sulphates        : num  0.32 0.46 0.42 0.54 0.34 0.43 0.47 0.47 0.47 0.39 ...
$ alcohol          : num  10.4 8.9 10.1 11.2 10.9 ...
$ quality          : int   5 6 6 4 6 6 6 6 6 7 ...

```

Figura 3: Panoramica tipi di variabile

Essendo le variabili di tipo numerico è possibile estrarre numerose statistiche riguardo la loro distribuzione tra cui minimi, massimi, somma, mediana, media, varianza e deviazione standard. Di seguito un sommario del loro calcolo:

	acidi_non_volatili	acidi_volatili	acido_citrico	zuccheri_residui	cloruri
nbr.val	4.898000e+03	4898.000000	4898.000000	4.898000e+03	4.898000e+03
nbr.null	0.000000e+00	0.000000	19.000000	0.000000e+00	0.000000e+00
nbr.na	0.000000e+00	0.000000	0.000000	0.000000e+00	0.000000e+00
min	3.800000e+00	0.080000	0.000000	6.000000e-01	9.000000e-03
max	1.420000e+01	1.100000	1.660000	6.580000e+01	3.460000e-01
range	1.040000e+01	1.020000	1.660000	6.520000e+01	3.370000e-01
sum	3.357475e+04	1362.825000	1636.870000	3.130515e+04	2.24193e+02
median	6.800000e+00	0.260000	0.320000	5.200000e+00	4.300000e-02
mean	6.854788e+00	0.2782411	0.3341915	6.391415e+00	4.57724e-02
SE.mean	1.205770e-02	0.0014402	0.0017292	7.247280e-02	3.122000e-04
CI.mean.0.95	2.363850e-02	0.0028235	0.0033900	1.420791e-01	6.120000e-04
var	7.121136e-01	0.0101595	0.0146458	2.572577e+01	4.773000e-04
std.dev	8.438682e-01	0.1007945	0.1210198	5.072058e+00	2.184800e-02
coef.var	1.231064e-01	0.3622561	0.3621271	7.935736e-01	4.77318e-01

	anidride_solforosa_libera	anidride_solforosa_totale	densita	ph	solfat	alcol	qualita
nbr.val	4.898000e+03	4.898000e+03	4898.000000	4.898000e+03	4898.000000	4.898000e+03	4.898000e+03
nbr.null	0.000000e+00	0.000000e+00	0.000000	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
nbr.na	0.000000e+00	0.000000e+00	0.000000	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
min	2.000000e+00	9.000000e+00	0.9871100	2.720000e+00	0.2200000	8.000000e+00	3.000000e+00
max	2.890000e+02	4.400000e+02	1.0389800	3.820000e+00	1.0800000	1.420000e+01	9.000000e+00
range	2.870000e+02	4.310000e+02	0.0518700	1.100000e+00	0.8600000	6.200000e+00	6.000000e+00
sum	1.729390e+05	6.776905e+05	4868.7460900	1.561613e+04	2399.2700000	5.149888e+04	2.879000e+04
median	3.400000e+01	1.340000e+02	0.9937400	3.180000e+00	0.4700000	1.040000e+01	6.000000e+00
mean	3.530808e+01	1.383607e+02	0.9940274	3.188267e+00	0.4898469	1.051427e+01	5.877909e+00
SE.mean	2.430087e-01	6.072391e-01	0.0000427	2.157600e-03	0.0016307	1.758390e-02	1.265460e-02
CI.mean.0.95	4.764061e-01	1.190461e+00	0.0000838	4.229800e-03	0.0031969	3.447230e-02	2.480860e-02
var	2.892427e+02	1.806085e+03	0.0000089	2.280120e-02	0.0130247	1.514427e+00	7.843557e-01
std.dev	1.700714e+01	4.249806e+01	0.0029909	1.510006e-01	0.1141258	1.230621e+00	8.856386e-01
coef.var	4.816783e-01	3.071543e-01	0.0030089	4.736130e-02	0.2329827	1.170429e-01	1.506724e-01

Figura 4: Calcolo delle principali statistiche riguardo la distribuzione dei valori delle variabili

Da una prima analisi delle statistiche è possibile notare come la qualità presenti una media di 5.87, molto vicino alla mediana di 6 facendo notare come la distribuzione dei vini sia centrata molto su una qualità di valore 6. Varianza e deviazione standard delle variabili risultano molto contenute, rispecchiando una minima dispersione dei valori.

2.3 Descrizione delle distribuzioni delle variabili

Per effettuare un'analisi approfondita delle variabili è possibile rappresentarne le distribuzioni attraverso istogrammi di occorrenza, box plots, q-q plots e densità.

2.3.1 Qualità

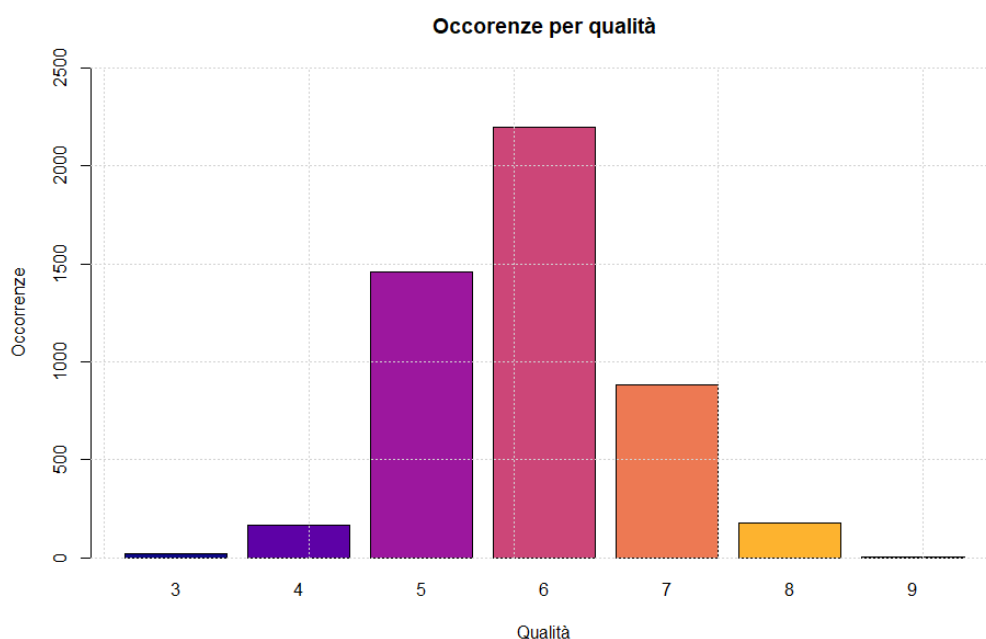


Figura 5: Istogramma delle occorrenze di vini rispetto alla loro qualità

Come si denota dall'istogramma delle occorrenze dei vini in base alla loro qualità, la presenza di vini di qualità 6 è predominante, mentre la loro distribuzione per valori minori e maggiori di 6 è pressoché equivalente, questo consentirà successivamente la creazione di una feature aggiuntiva riguardo il gusto del vino: utilizzando la qualità 6 per denotare un gusto normale, una qualità maggiore di 6 per denotare un buon gusto e una qualità minore di 6 per denotare un cattivo gusto.

2.3.2 Acidi non volatili

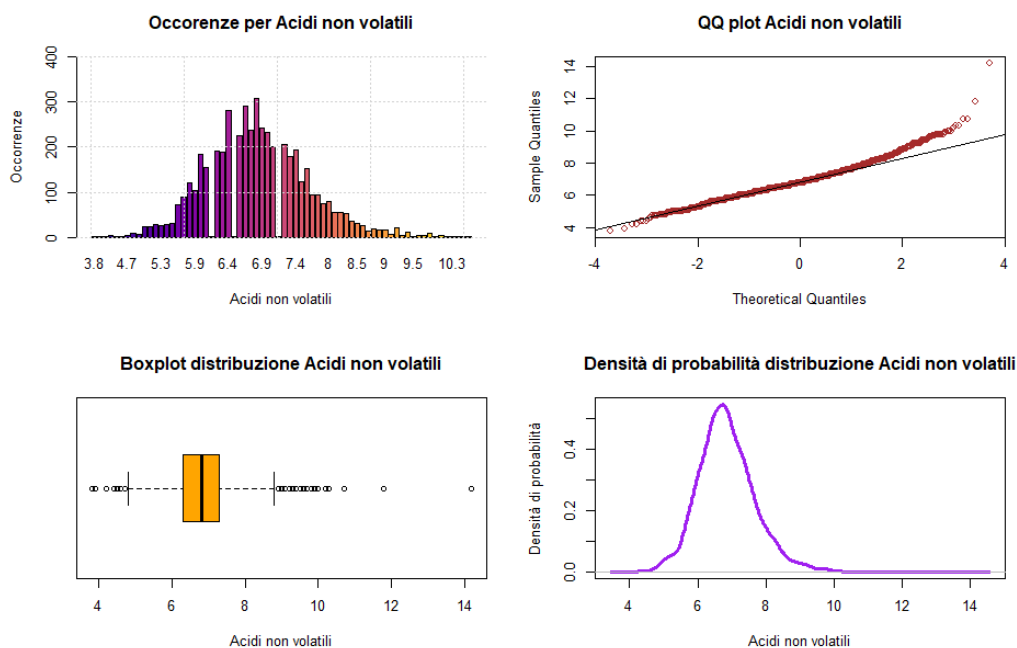


Figura 6: Plots per la descrizione della distribuzione della variabile di Acidi non volatili

Dall'insieme dei grafici relativi alla distribuzione degli acidi non volatili è possibile concludere che la maggior parte dei valori seguano una distribuzione normale, con l'eccezione di numerosi outliers visualizzabili nel boxplot nel range dal terzo all'ultimo quartile.

2.3.3 Acidi volatili

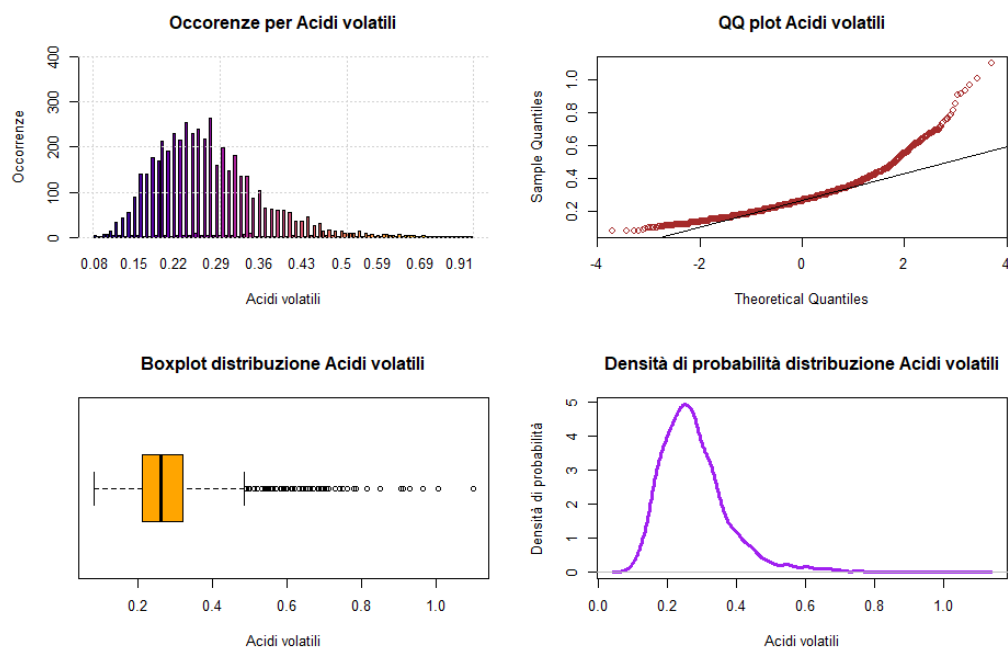


Figura 7: Plots per la descrizione della distribuzione della variabile di Acidi volatili

A differenza degli acidi non volatili, dall'insieme dei grafici relativi alla distribuzione degli acidi volatili è possibile concludere che i valori si discostano maggiormente da una distribuzione normale. Infatti anche dal boxplot è possibile notare una grossa presenza di outliers per alti valori di acidi volatili.

2.3.4 Acido citrico

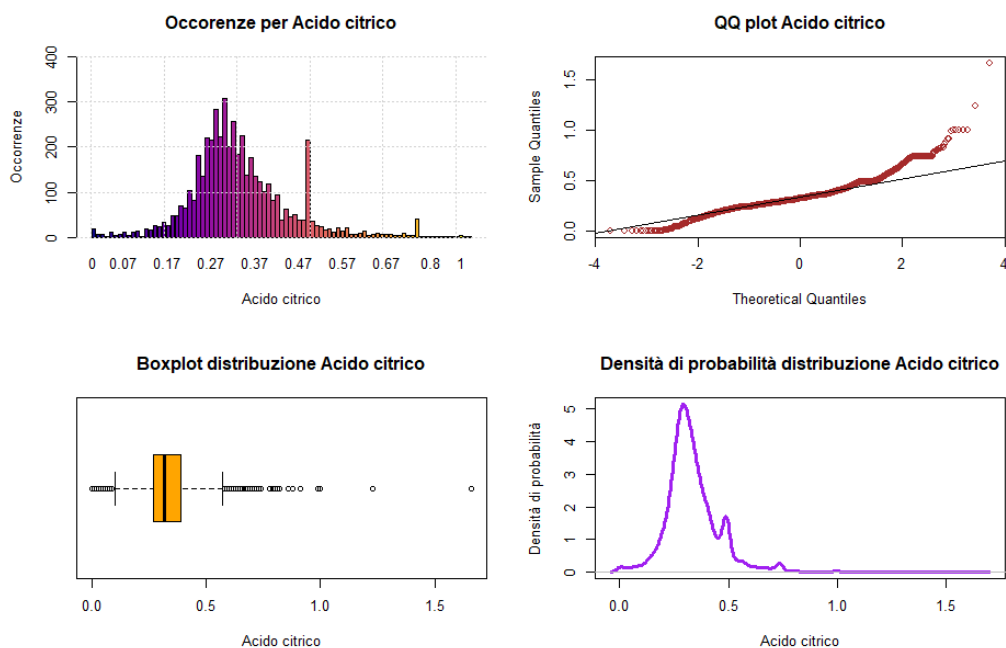


Figura 8: Plots per la descrizione della distribuzione della variabile di Acido citrico

L'insieme dei grafici relativi alla distribuzione dei valori di acido citrico ne denota una certa irregolarità. Questa considerazione può essere fatta dall'osservazione di numerosi outliers, sia per bassi che alti valori, della distribuzione dei valori di acido citrico. Spicca un valore su tutti in corrispondenza di 2 g/L.

2.3.5 Zuccheri residui

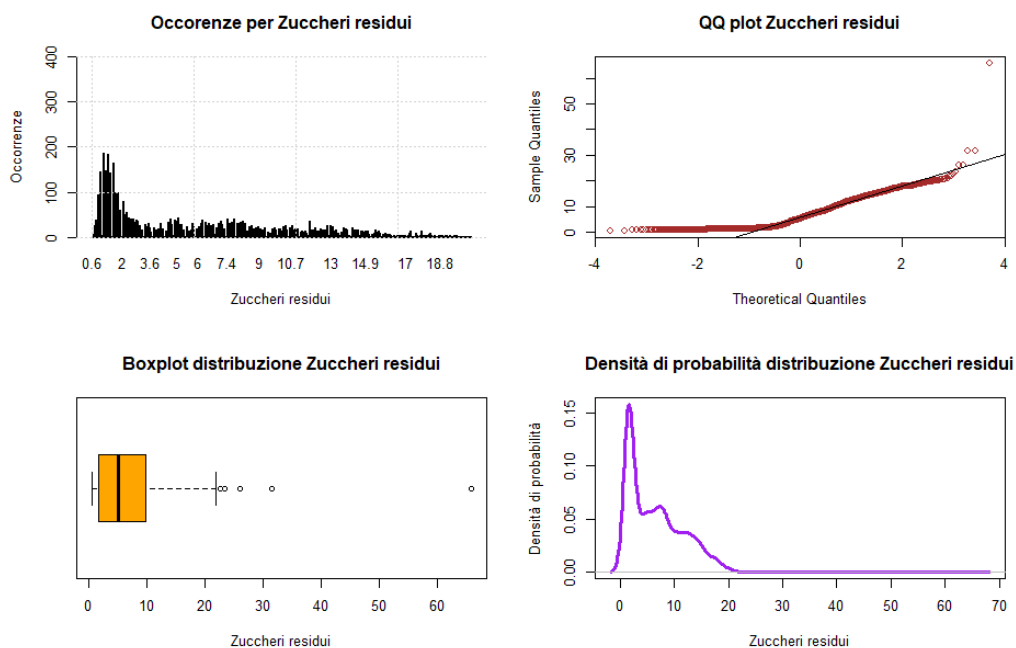


Figura 9: Plots per la descrizione della distribuzione della variabile di Zuccheri residui

Così come i valori di acido citrico anche l'insieme dei grafici relativi alla distribuzione dei valori di zuccheri residui ne denota una certa irregolarità. Infatti è possibile notare dal boxplot e dal Q-Q plot come i valori siano molto più concentrati sui minimi. Si segnala anche la presenza di un outlier con un valore di circa 70 g/L, un'enormità.

2.3.6 Cloruri

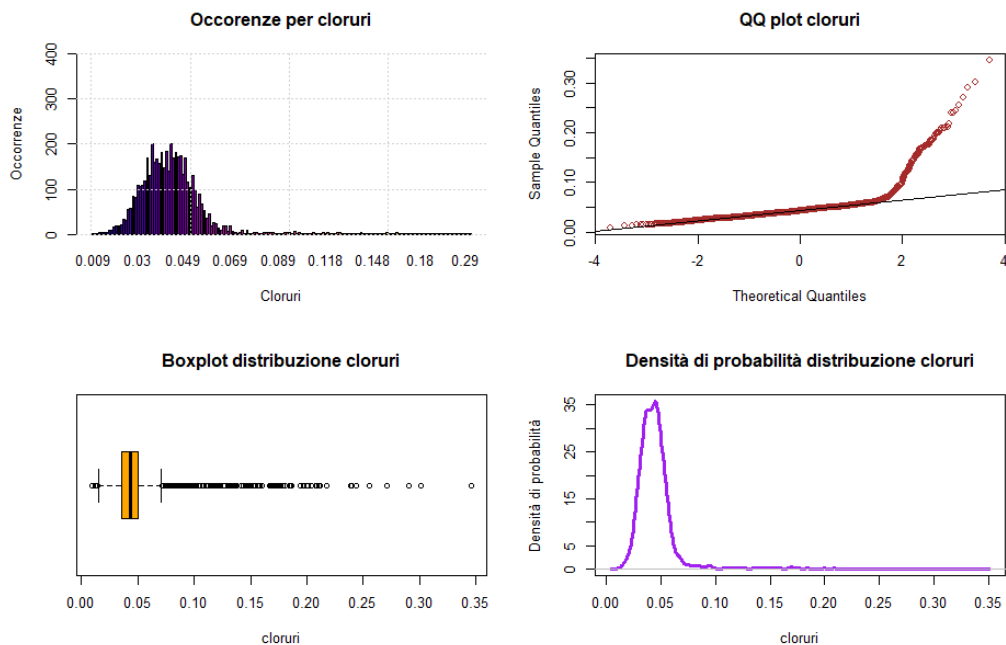


Figura 10: Plots per la descrizione della distribuzione della variabile di Cloruri

Dall'insieme dei grafici relativi alla distribuzione dei cloruri è possibile concludere che la maggior parte dei valori seguano una distribuzione normale per bassi valori di cloruro. Invece si può notare come per alti valori cloruro, la distribuzione si discosta molto da una normale e appaiono quindi molti outliers.

2.3.7 Anidride Solforosa libera

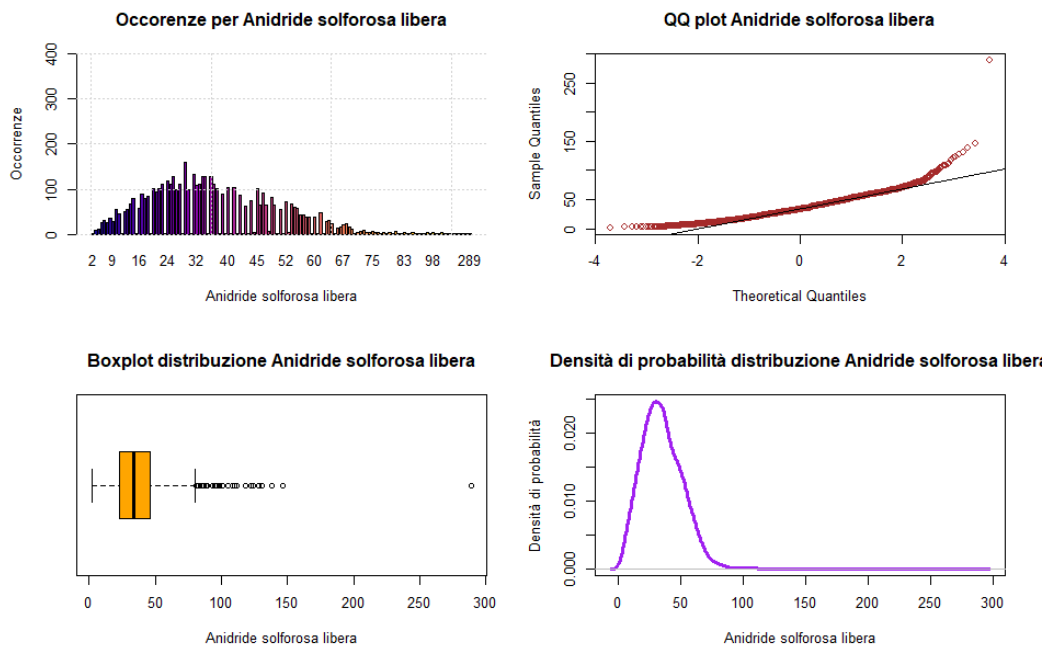


Figura 11: Plots per la descrizione della distribuzione della variabile di Anidride Solforosa libera

Dall'insieme dei grafici relativi alla distribuzione della variabile di Anidride Solforosa libera è possibile concludere che la maggior parte dei valori seguano una distribuzione normale, con l'eccezione di alcuni outlier per alti valori. Si può notare anche la presenza di un outlier in corrispondenza del valore di 300 mg/L. Molto al di sopra del limite consentito dalla legge europea (210 mg/l nei vini bianchi e rosati).

2.3.8 Anidride Solforosa totale

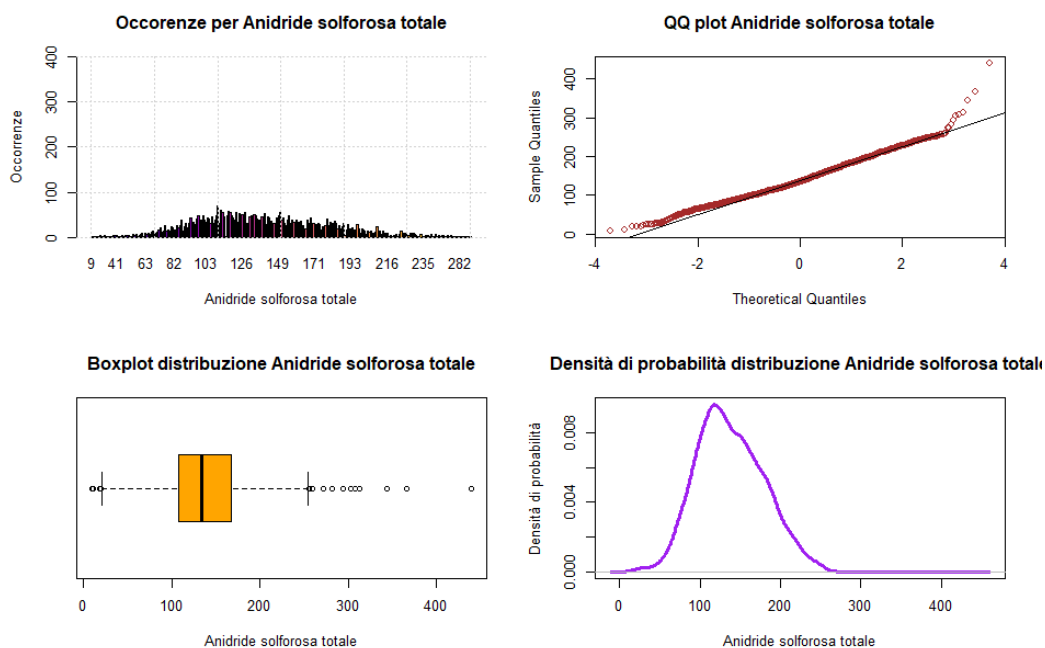


Figura 12: Plots per la descrizione della distribuzione della variabile di Anidride Solforosa totale

Così come l'insieme dei grafici relativi alla distribuzione della variabile di Anidride Solforosa libera anche i grafici relativi alla distribuzione della variabile di Anidride Solforosa totale non si discostano molto da una distribuzione normale. Si può notare inoltre la minor presenza di outliers rispetto alla variabile precedente.

2.3.9 Densità

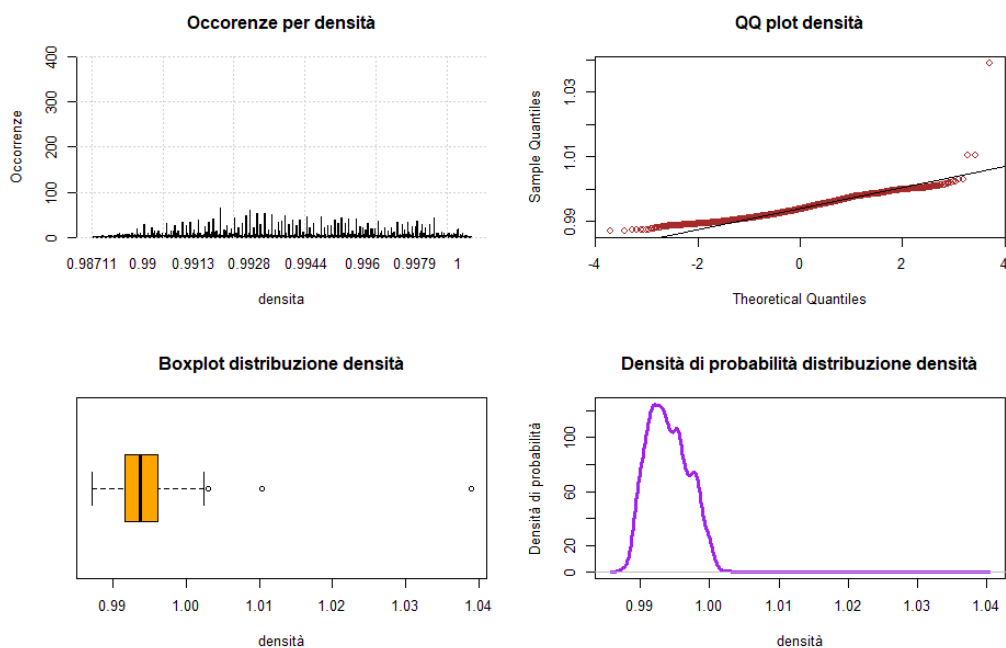


Figura 13: Plots per la descrizione della distribuzione della variabile di Densità

Dall'insieme dei grafici relativi alla distribuzione della variabile di Densità è possibile concludere che la quasi totalità di valori seguano una distribuzione normale, con l'eccezione di alcuni outlier per alti valori. In particolare si può notare un outlier in corrispondenza del valore di 1.04 di densità.

2.3.10 pH

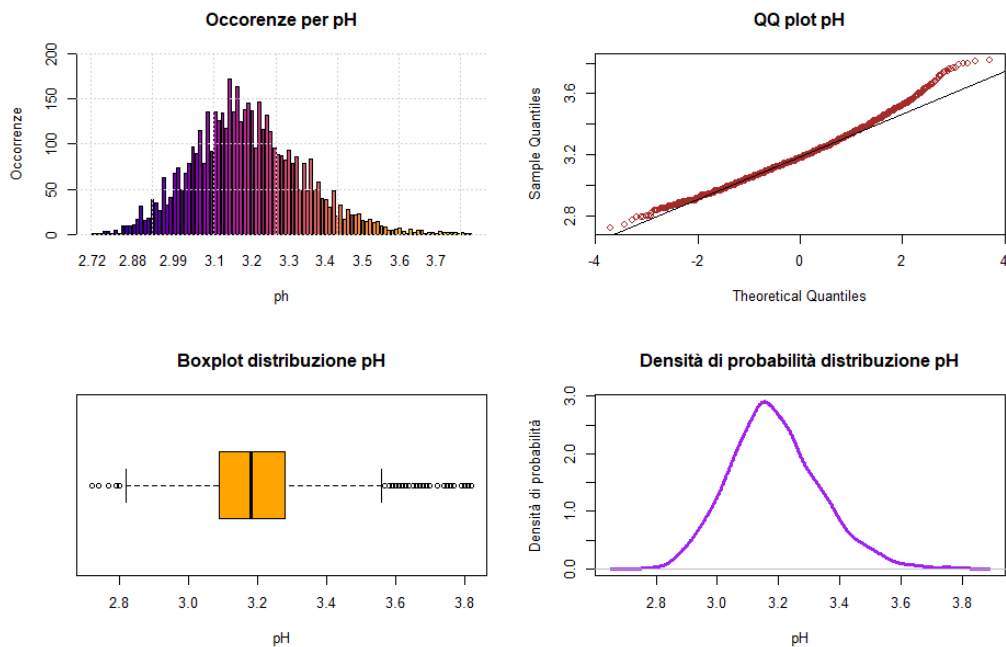


Figura 14: Plots per la descrizione della distribuzione della variabile di pH

Così come l'insieme dei grafici relativi alla distribuzione della variabile di densità anche i grafici relativi alla distribuzione della variabile di pH non si discostano per nulla da una distribuzione normale. Si può comunque notare la presenza di diversi outlier più concentrati su alti valori di pH

2.3.11 Solfati

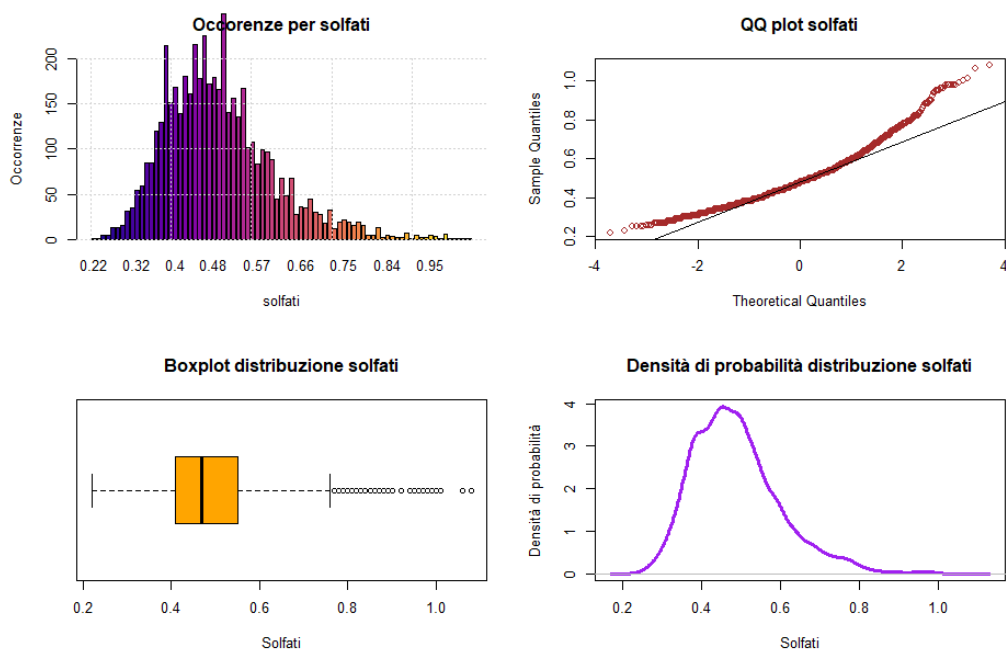


Figura 15: Plots per la descrizione della distribuzione della variabile di Solfati

Anche la distribuzione della variabile relativa ai solfati non si discosta molto da una normale, anche se meno delle due variabili precedenti (ph e densità). Ci sono inoltre numerosi outliers per alti valori di solfati.

2.3.12 Percentuale di Alcol

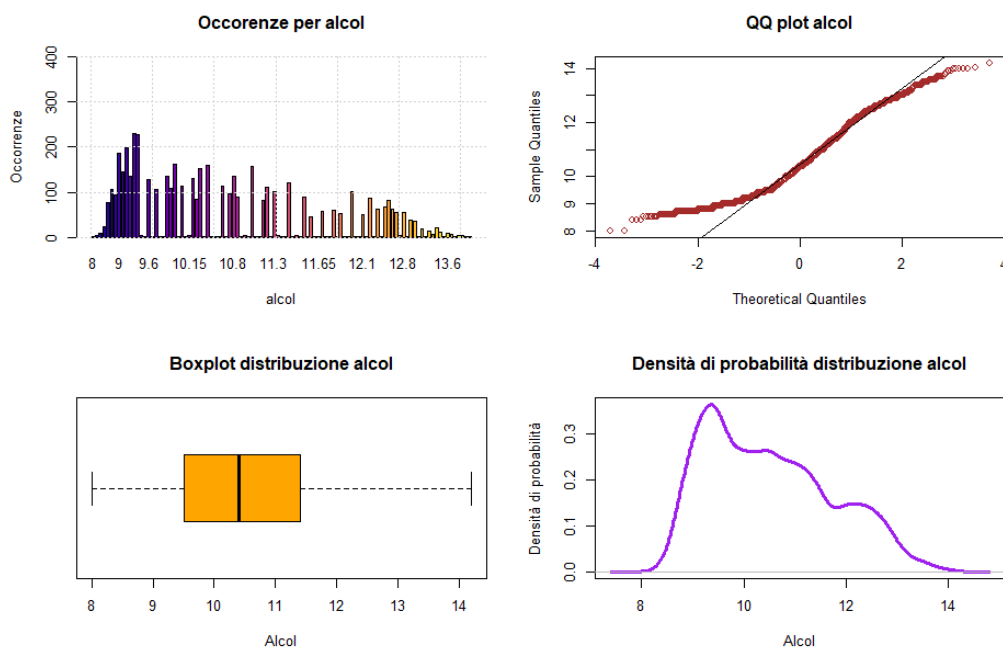


Figura 16: Plots per la descrizione della distribuzione della variabile di Percentuale di Alcol

La distribuzione della variabile di percentuale alcolica è discretamente regolare, infatti non presenta alcun outlier. Si può notare però dal Q-Q plot relativo come non sia distribuita propriamente secondo una normale.

Conclusion Non presentando casi eccezionali di outliers o distribuzioni non regolari è possibile proseguire con l'analisi e i task di regressione e classificazione regolarmente senza intaccare l'integrità del dataset.

2.4 Correlazione tra variabili

Utile ai fini dell'analisi è lo studio della matrice di correlazione tra le variabili del dataset per poter meglio comprendere da che parametri la qualità sia maggiormente "influenzata". Di seguito viene riportata una rappresentazione della matrice di correlazione tra le variabili.

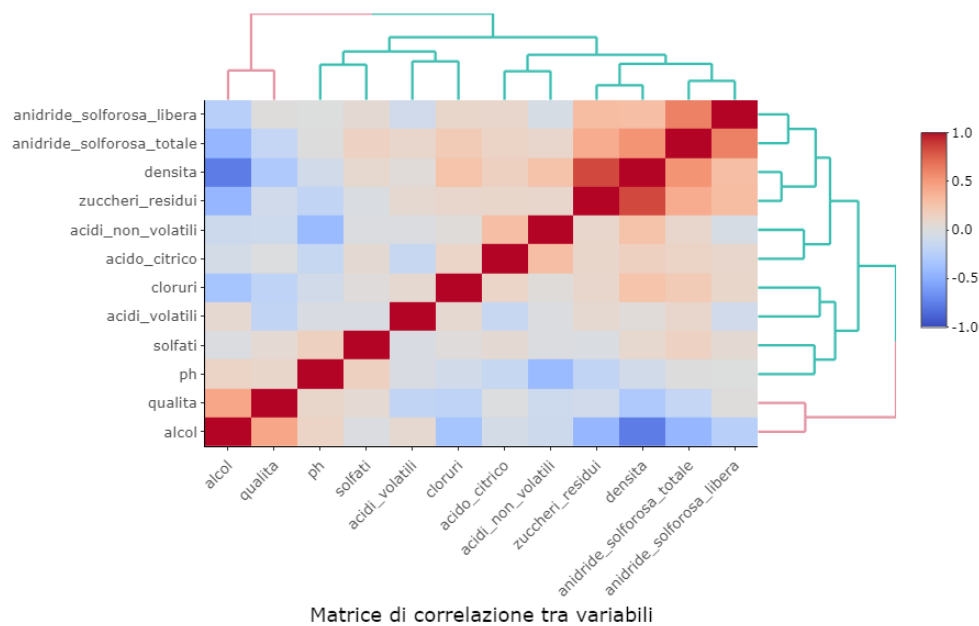


Figura 17: Matrice di correlazione tra le variabili

Da una generica analisi della matrice di correlazione si può notare come la qualità del vino sia discretamente correlata (direttamente) con la percentuale alcolica e (inversamente) con la densità del vino. Ci si aspetta quindi che nei task di regressione e di classificazione vengano considerati con maggiore importanza questi due parametri.

3 Regression e Classificazione

3.1 Organizzazione di training e test set

Per l'applicazione di metodologie di predizione e classificazione e la loro valutazione il dataset è stato così suddiviso:

- 85% Training: 4163 Entries
- 15% Test: 735 Entries

3.2 Regressione

3.2.1 Regressione Lineare Multipla

Il task prevede l'implementazione di modelli di regressione per la predizione della **qualità** dei vini in base alle loro proprietà chimiche. La formula iniziale che contiene tutte le variabili a disposizione prevede l'utilizzo della *qualità* come variabile dipendente e la totalità delle altre variabili come variabili indipendenti.

```
call:
lm(formula = init_formula, data = training_values)

Residuals:
    Min       1Q   Median       3Q      Max
-3.4717 -0.4901 -0.0344  0.4563  3.1588

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.410e+02   1.988e+01   7.091 1.56e-12 ***
acidi_non_volatili  4.930e-02   2.239e-02    2.202  0.0277 *
acidi_volatili    -1.839e+00   1.219e-01  -15.084 < 2e-16 ***
acido_citrico     2.715e-03   1.040e-01    0.026  0.9792
zuccheri_residui  7.665e-02   8.014e-03   9.565 < 2e-16 ***
cloruri          -2.574e-01   5.873e-01   -0.438  0.6611
anidride_solforosa_libera  4.342e-03   9.244e-04   4.698 2.72e-06 ***
anidride_solforosa_totale -7.132e-05   4.087e-04  -0.175  0.8615
densita         -1.409e+02   2.017e+01  -6.986 3.28e-12 ***
ph              6.483e-01   1.133e-01    5.724 1.11e-08 ***
solfati         5.899e-01   1.088e-01    5.423 6.18e-08 ***
alcol           2.059e-01   2.567e-02    8.022 1.35e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7517 on 4151 degrees of freedom
Multiple R-squared:  0.2823,    Adjusted R-squared:  0.2804
F-statistic: 148.4 on 11 and 4151 DF,  p-value: < 2.2e-16
```

Figura 18: Modello di regressione lineare multipla per la predizione della qualità del vino utilizzando tutte le variabili a disposizione

Come si può notare ci sono delle variabili che non presentano una significatività sufficiente. Queste variabili vengono eliminate dal modello di regressione tramite l'utilizzo della tecnica di backward elimination. Un altro

aspetto rilevante è il basso valore di adjusted R squared che denota una bassa efficacia del modello di predizione. Questo viene confermato anche dalla statistica relativa al Residual Standard Error, pari a 0.7517.

```
Call:
lm(formula = improved_formula, data = training_values)

Residuals:
    Min       1Q   Median       3Q      Max
-3.4687 -0.4916 -0.0370  0.4554  3.1609

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.430e+02  1.918e+01   7.457 1.07e-13 ***
acidi_volatili -1.849e+00  1.170e-01 -15.805 < 2e-16 ***
acidi_non_volatili  5.097e-02  2.194e-02   2.323  0.0202 *
zuccheri_residui  7.753e-02  7.773e-03   9.974 < 2e-16 ***
anidride_solforosa_libera  4.230e-03  7.465e-04   5.666 1.56e-08 ***
densita      -1.430e+02  1.944e+01  -7.357 2.26e-13 ***
ph           6.568e-01  1.112e-01   5.907 3.76e-09 ***
solfati       5.902e-01  1.083e-01   5.449 5.35e-08 ***
alcol        2.059e-01  2.551e-02   8.069 9.20e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7514 on 4154 degrees of freedom
Multiple R-squared:  0.2822,    Adjusted R-squared:  0.2808
F-statistic: 204.2 on 8 and 4154 DF,  p-value: < 2.2e-16
```

Figura 19: Modello di regressione lineare multipla per la predizione della qualità del vino utilizzando le variabili più significative

Come si può notare il valore di adjusted R squared è aumentato di 0.0004, un incremento poco significativo per concludere di poter predire correttamente la qualità di un vino con un modello di regressione di questo tipo. Questo è confermato anche dalla stima del MAE (Mean Absoulte Error) che passa da un valore iniziale di 0.5935 a un valore di 0.5939.

3.2.2 Albero di regressione semplice

Per implementare il task di regressione è possibile anche modellare un albero di decisione sulle variabili del dataset, ottenendo il modello seguente

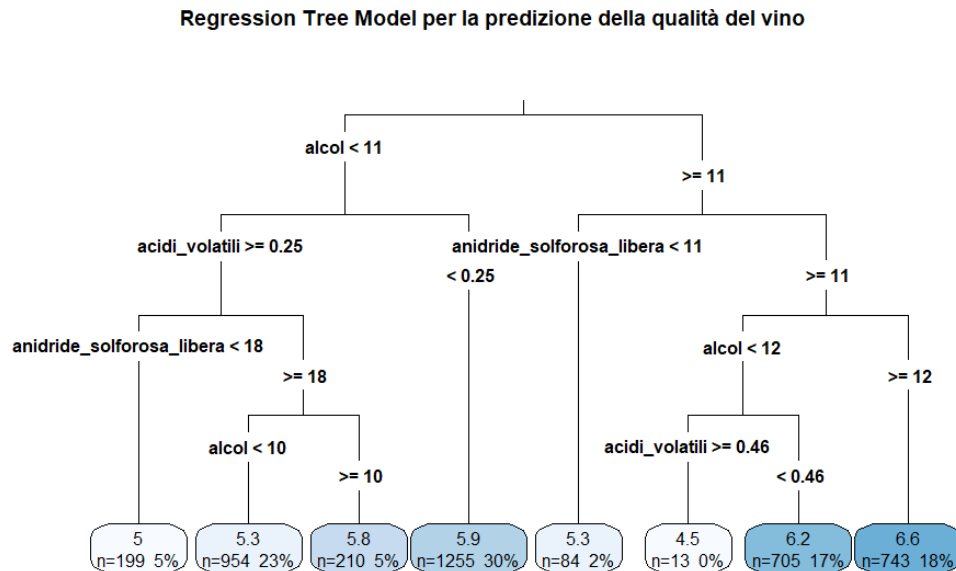


Figura 20: Modello di regressione ad albero per la predizione della qualità del vino utilizzando tutte le variabili

Questo modello di regressione produce un MAE pari a 0.6103, un valore ancora peggiore rispetto ai modelli di regressione lineare multipla precedenti e quindi insufficiente.

3.2.3 M5P

M5P è una ristrutturazione dell'algoritmo M5 di Quinlan, dove viene combinato un albero decisionale convenzionale con la possibilità di aggiungere funzioni di regressione lineare nei nodi. Questo modello di regressione ha stimato i valori di qualità del test set ottenendo un MAE di 0.3887, nettamente inferiore ai modelli precedenti e accettabile per un modello di predizione.

Conclusione L'unico modello accettabile per predire la qualità del vino in base alle sue caratteristiche chimiche è risultato essere l'M5P. Anche se non ottimale, risulta comunque essere un discreto strumento da utilizzare. Di seguito il sommario dei MAE dei vari modelli di regressione.

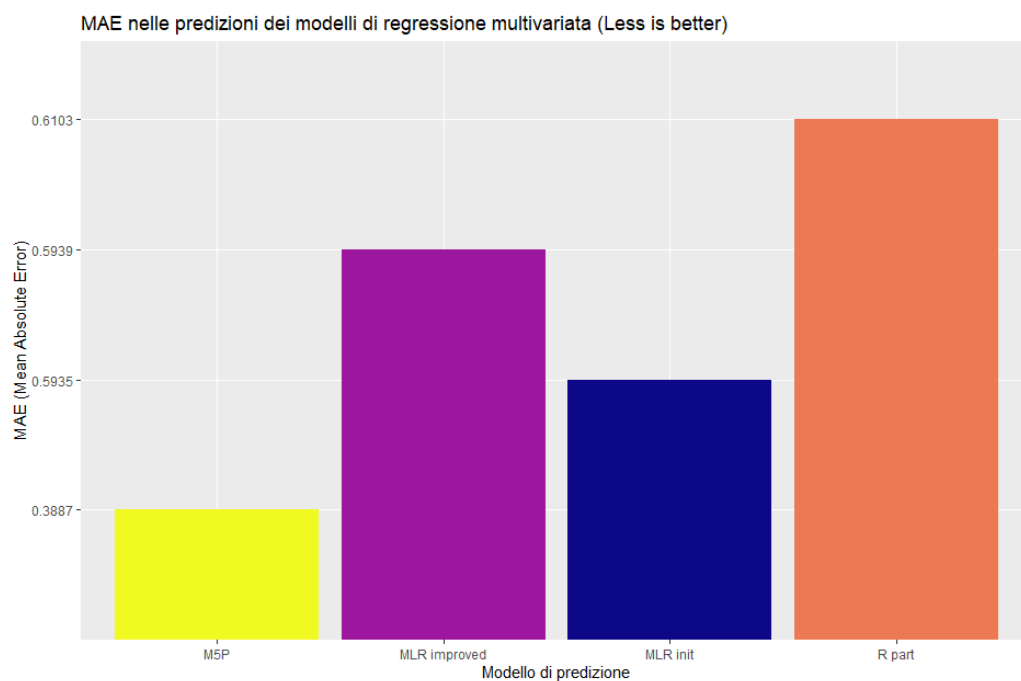
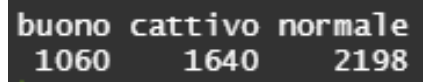


Figura 21: Sommario dei MAE dei modelli di regressione

3.3 Classificazione

3.3.1 Introduzione

Per effettuare una classificazione affidabile della qualità dei vini è stato deciso di suddividere il dataset utilizzando una feature aggiuntiva (gusto), che raggruppa la qualità dei vini in tre macrogruppi: di buon gusto (qualità > 6), di normale gusto (qualità $= 6$) e di cattivo gusto (qualità < 6). Di seguito il numero di occorrenze per gruppo:



buono	cattivo	normale
1060	1640	2198

Figura 22: Occorrenze di vino per gusto

3.3.2 Albero di classificazione semplice

Per una semplice classificazione della qualità del vino è possibile utilizzare ancora un modello decisionale ad albero. Di seguito la sua applicazione: Dopo la predizione della classificazione dei vini e il confronto con i dati reali

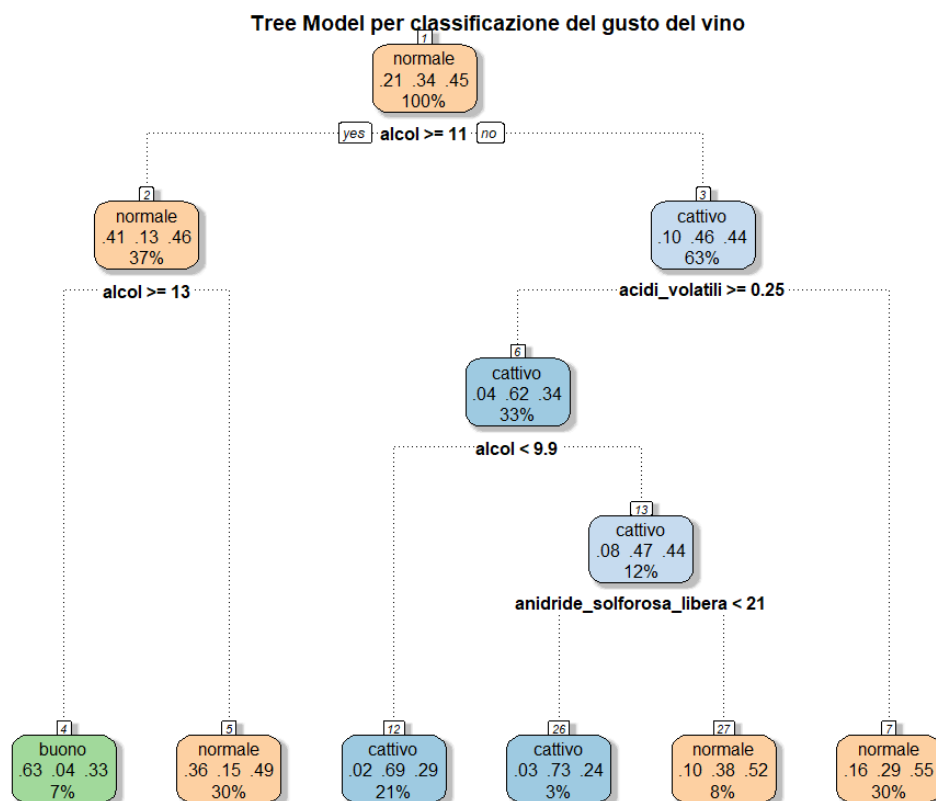


Figura 23: Albero di classificazione dei vini

è stata ottenuta la seguente matrice di confusione

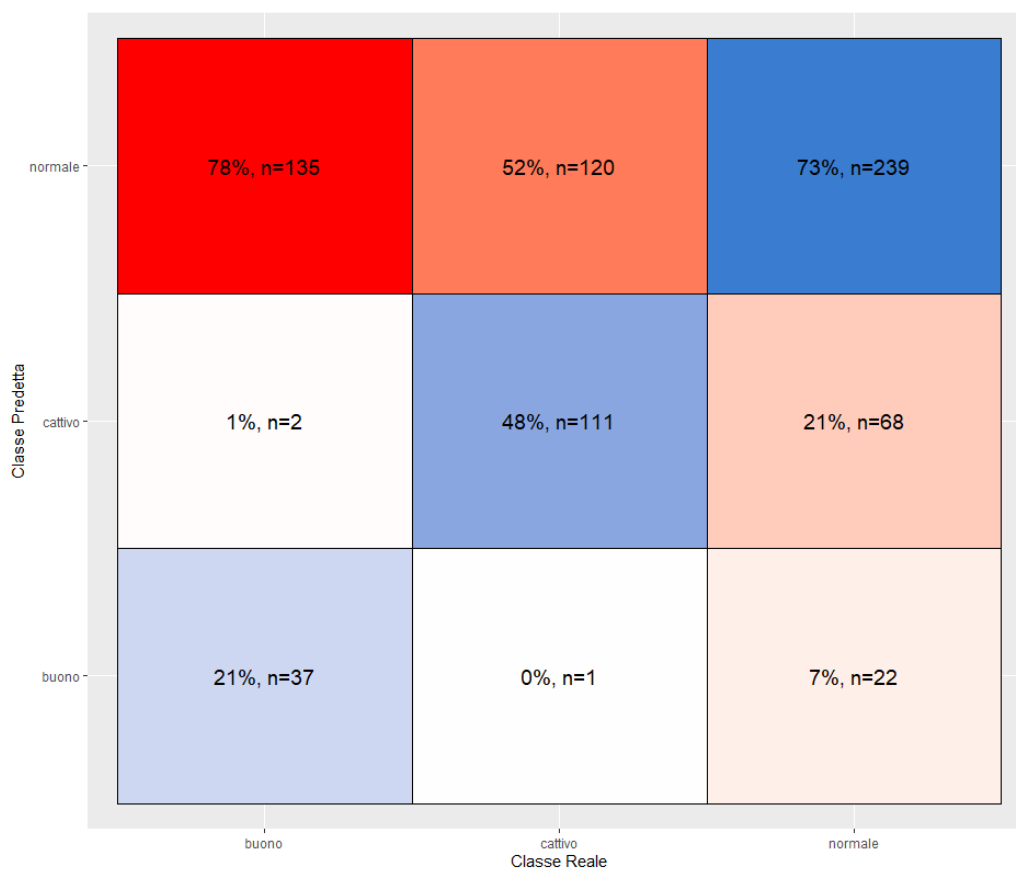


Figura 24: Matrice di confusione per il modello di classificazione ad albero semplice

La predizione tramite quello modello ha ottenuto un'accuracy pari al 52.65%, un valore poco accettabile.

3.3.3 Random Forest

Un altro metodo di classificazione implementabile è quello Random Forest. Partendo da un modello che fa uso di 100 alberi di decisione viene ottenuto il seguente risultato:

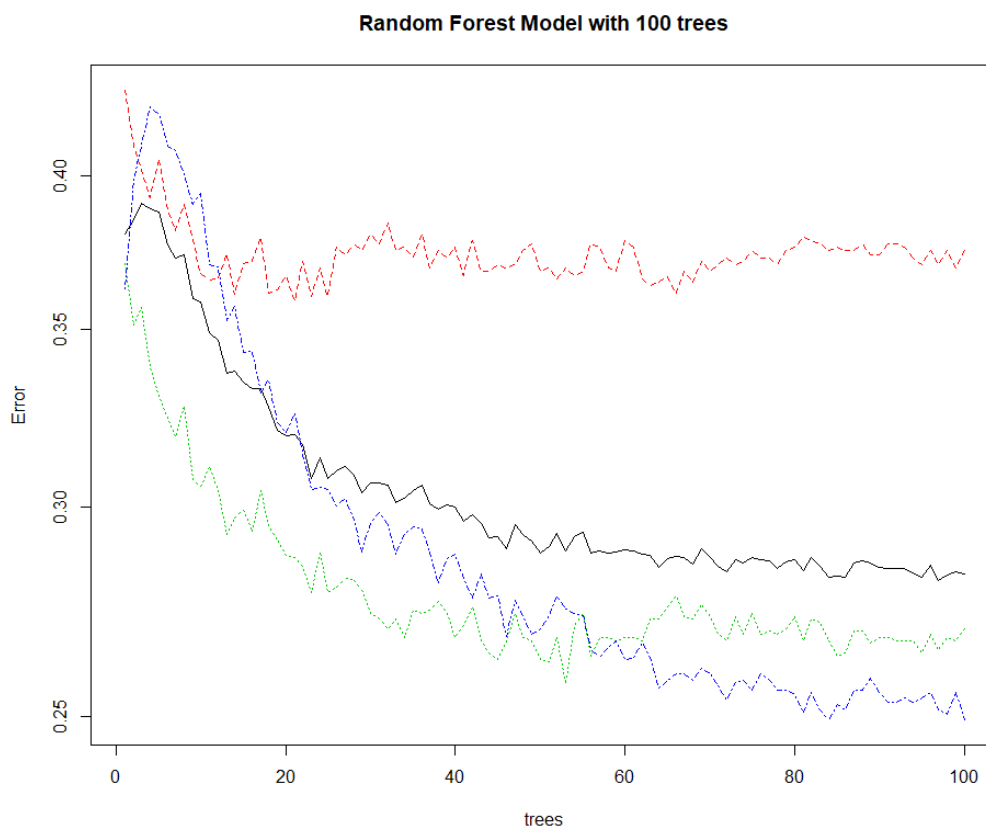


Figura 25: Modello Random Forest con 100 alberi di decisione

E la corrispondente matrice di confusione:

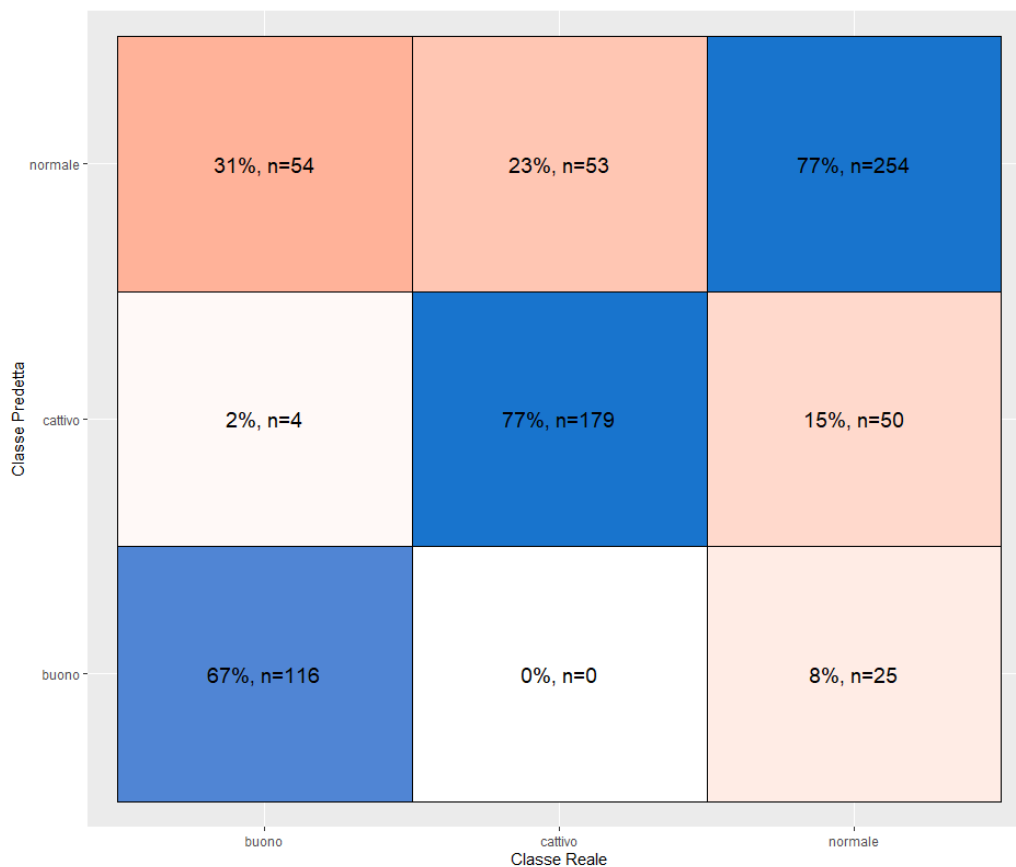


Figura 26: Matrice di confusione random forest con 100 alberi

La predizione Random Forest con 100 alberi di decisione ha prodotto un'accuracy del 74.69% un valore più che accettabile e di gran lunga superiore rispetto a quello prodotto dall'utilizzo di un albero di decisione semplice. Aumentando il numero di alberi prodotti da Random Forest si raggiungono risultati pressoché simili. Di seguito le matrici di confusione associate:

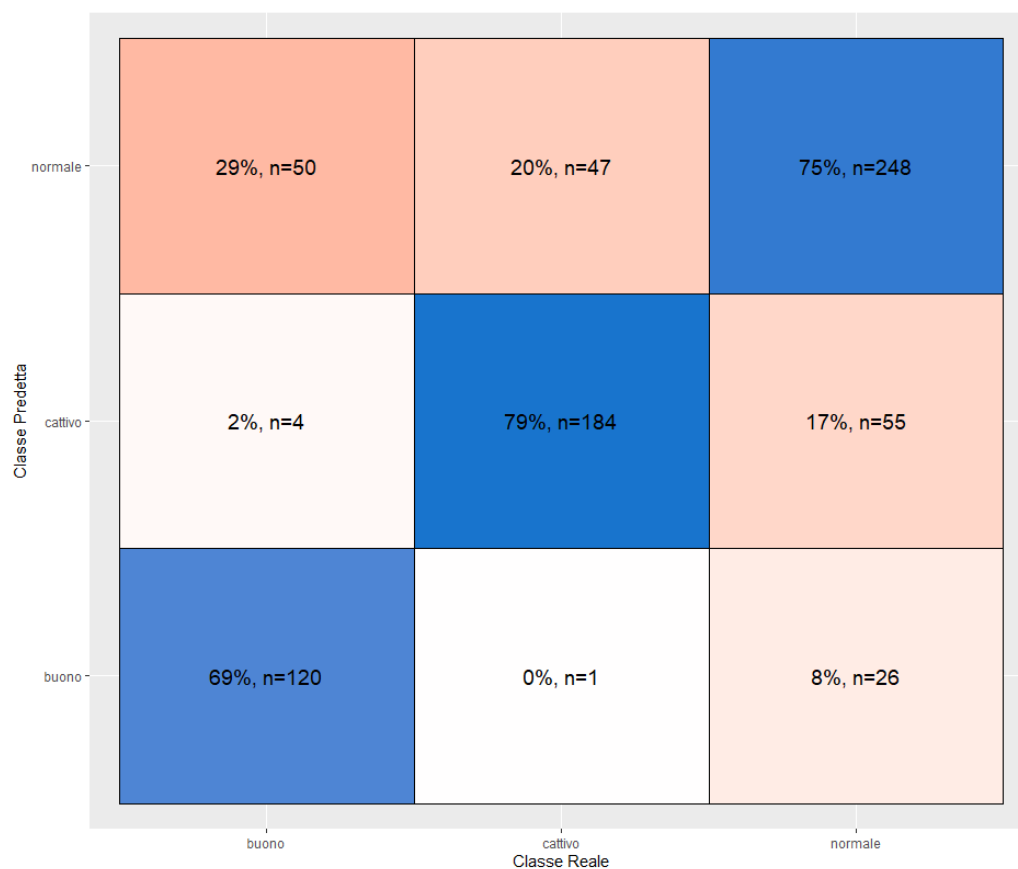


Figura 27: Matrice di confusione random forest con 200 alberi

Precisione: 75.1%

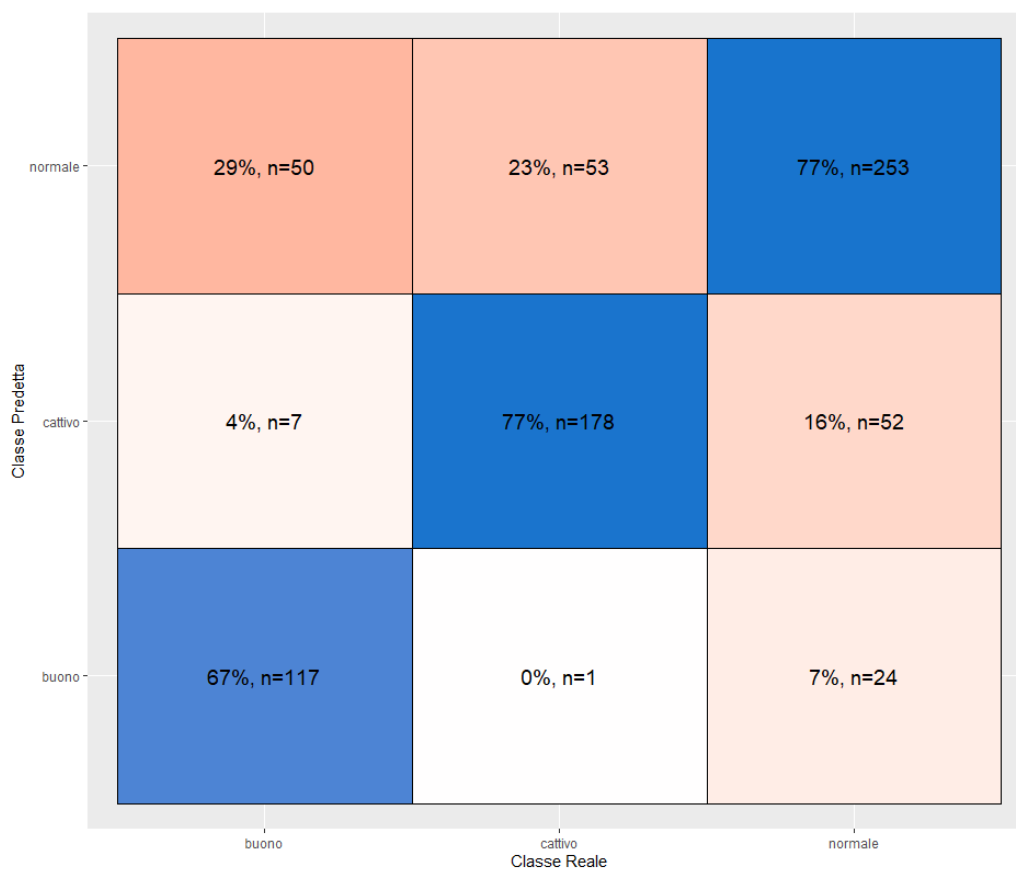


Figura 28: Matrice di confusione random forest con 300 alberi

Precisione: 74.56%

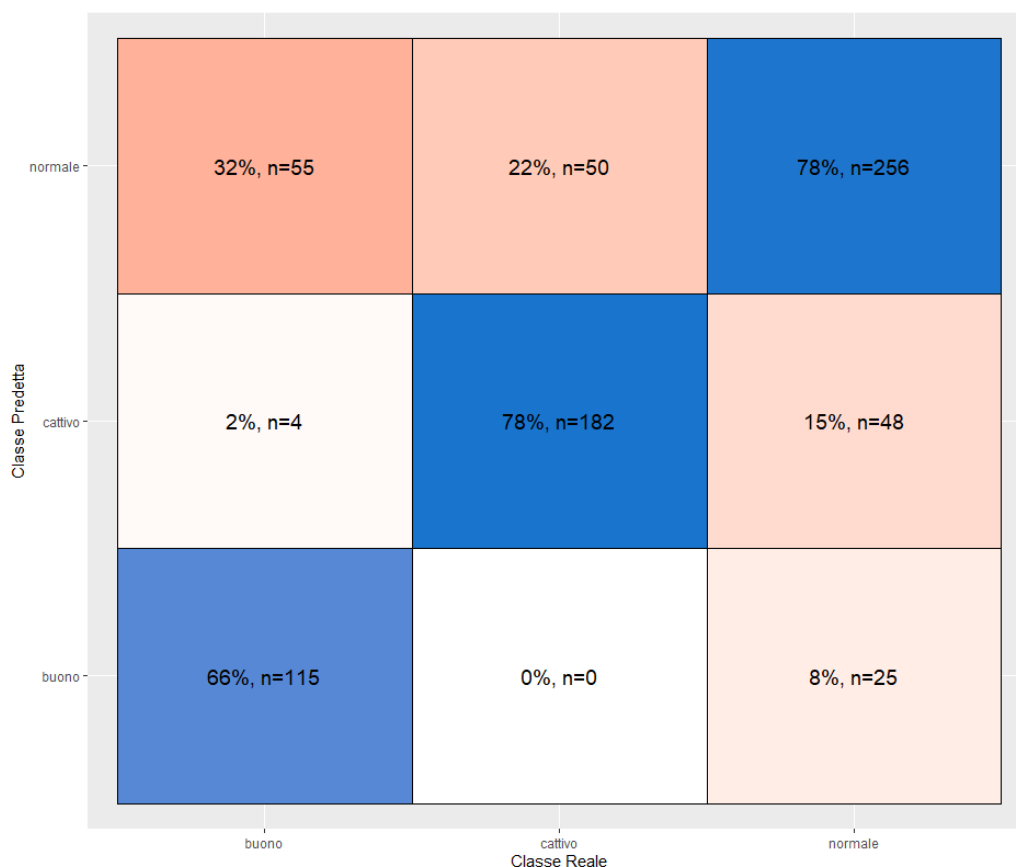


Figura 29: Matrice di confusione random forest con 400 alberi

Precisione: 75.24%

Conclusioni Dai modelli di classificazione presentati si evince che il migliore sia di gran lunga il random forest. Il numero di alberi utilizzati non influisce di molto sull'accuratezza dei risultati finali, quindi per avere migliori prestazioni è raccomandabile l'utilizzo del modello con 100 alberi. Di seguito il sommario dei risultati di accuratezza dei vari modelli:

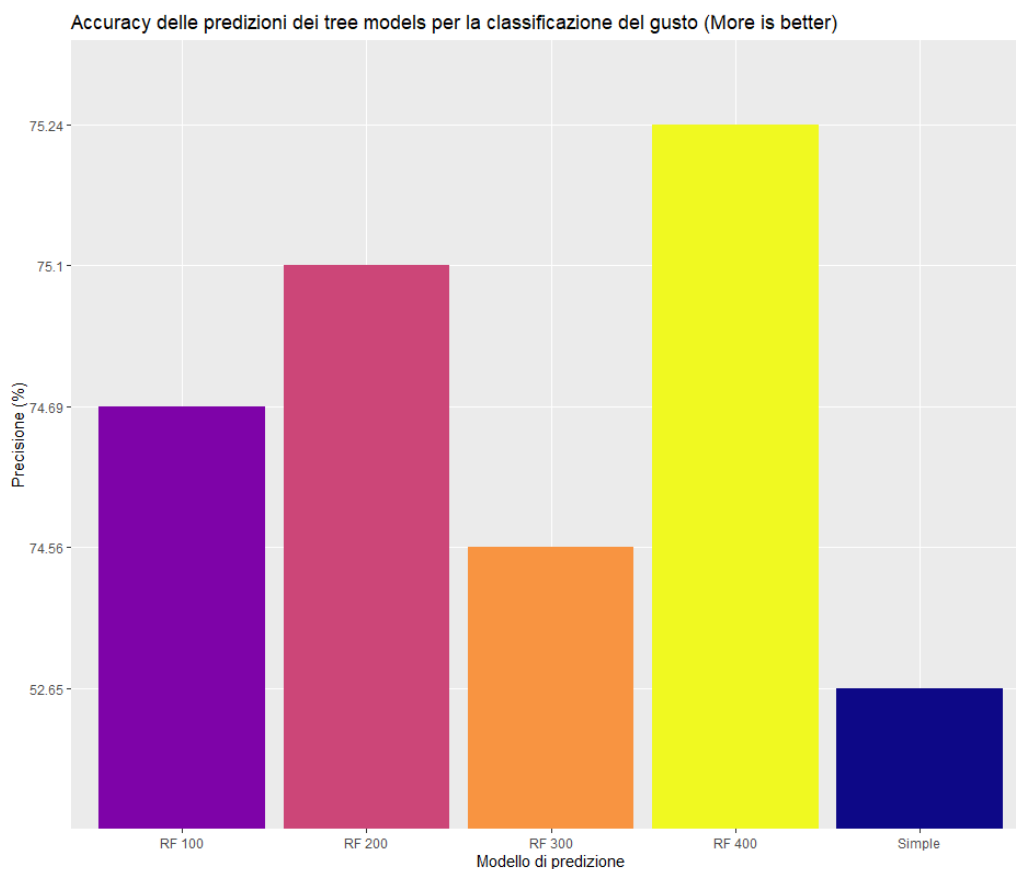


Figura 30: Accuratezza dei modelli di classificazione

4 Conclusioni

Dall'analisi dei dati prodotti dai modelli di regressione è possibile concludere che le proprietà chimiche in analisi non sono ottimali per una corretta stima della qualità del vino. È stato possibile però, dopo aver raggruppato i vini in base alla qualità in 3 macro gruppi, effettuare una classificazione con discreti risultati. In particolare mediante l'implementazione di Random Forest è stato possibile raggiungere un'accuratezza del 75%, questo denota come sia possibile stabilire il gusto di un vino particolarmente bene tra cattivo, normale e buono in base alle sue proprietà chimiche.