Università degli studi di Milano-Bicocca

A.A 2020/2021

Machine Learning

Progetto Machine Learning
Wine Quality - Prediction

Febbrario 2021

RAVIOTTA Benedetto 816362

SALAMONE Fabio 816297

Indice

1 Descrizione dominio e obiettivi

L'industria del vino bianco mostra una recente crescita esponenziale poiché il consumo sociale è in aumento. Al giorno d'oggi, gli operatori del settore utilizzano le certificazioni di qualità dei prodotti per promuovere i loro prodotti. Questo è un processo che richiede tempo e richiede la valutazione fornita da esperti umani, il che rende questo processo molto costoso. Inoltre, il prezzo del vino bianco dipende da un concetto piuttosto astratto di apprezzamento del vino da parte degli assaggiatori, l'opinione tra i quali può avere un alto grado di variabilità. Un altro fattore vitale nella certificazione del vino bianco e nella valutazione della qualità sono i test fisico-chimici, che sono basati su laboratorio e considerano fattori come l'acidità, il livello di pH, lo zucchero e altre proprietà chimiche. Il mercato del vino bianco sarebbe interessante se la qualità umana della degustazione potesse essere correlata alle proprietà chimiche del vino in modo che i processi di certificazione e valutazione della qualità e di garanzia siano più controllati. Questo progetto mira a determinare quali sono le caratteristiche per dare un'indicazione di qualità del vino bianco e sperimentare diversi metodi di classificazione per vedere quale produce la migliore precisione.

Obiettivi L'obiettivo stabilito per questo progetto è lo sviluppo di due modelli di apprendimento che abbiano lo scopo di classificare la qualità del vino bianco. I modelli di riferimento comprendono l'apprendimento attraverso le macchine a vettori di supporto e l'algoritmo random forest.

Fasi Inizialmente abbiamo importato il dataset in formato tabellare csv disponibile dal database UCI Machine Learning Repositor. Dopo di che abbiamo svolto un'analisi esplorativa da cui è possibile ricavare informazioni importanti e misurazioni necessarie per poter determinare l'andamento dei modelli di apprendimento. In seguito si svolge la fase di test di ogni modello, performando in tal modo le misurazioni delle performance su ciascuno di esso, ottenendo risultati sia grafici che tabellari.

1.1 Comprensione del dominio

L'analisi utilizza il dataset Red Wine Quality Data Set disponibile su UCI machine learning repository. Il dataset utilizzato ha le seguenti caratteristiche:

- Caratteristiche del dataset: **multivariato**
- Numero di istanze: **4898**
- Caratteristiche degli attributi: **Reali**

- Numero di attributi: **12**
- Valori nulli : **N/A**

Di seguito una breve descrizione degli attributi:

- **Acidità fissa** : sono acidi non volatili che non evaporano facilmente
- **Acidità volatile**: sono livelli di acido acetico nel vino che porta ad uno sgradevole sapore di aceto
- **Acido citrico**: funge da conservante per aumentare l'acidità (piccole quantità aggiungono freschezza e sapore ai vini)
- **Zuccheri residui**: è il quantità di zucchero rimanente dopo l'arresto della fermentazione. La chiave è avere un perfetto equilibrio tra - dolcezza e acidità (i vini > 45 g/l sono dolci)
- **Cloruri**: la quantità di sale nel vino
- **Anidride solforosa libera**: previene la crescita microbica e l'ossidazione del vino
- **Anidride solforosa totale**: è la quantità di SO2 legate e libere
- **Densità**: i vini più dolci hanno una densità maggiore
- **pH**: il livello di acidità
- **Solfati**: un additivo per vino che contribuisce ai livelli di SO2 e agisce come antimicrobico e antiossidante
- **Alcool**: quantità di alcol presente nel vino
- **Qualità**: A ogni vino è stato assegnato un punteggio di "qualità" compreso tra 0 e 10.

2 Analisi dataset

Ai fini di questo progetto, inizialmente l'output è stato convertito in un output ternario in cui ogni vino è "di Alta qualità" (un punteggio di 8 o superiore), "Media qualità" (un punteggio di 6 o 7) o "Bassa qualità" (un punteggio di 5 o inferiore). Non è stato necessario gestire alcun valore mancante all'interno del dataset.

Osservando la distribuzione della variabile *qualità*, si nota una distribuzione normale con valori medi di qualità 5 e 6. Inoltre non sono presenti istanze di vini con qualità 1, 2, e 10.
In figura 1 è mostra la distribuzione della variabile target qualità.

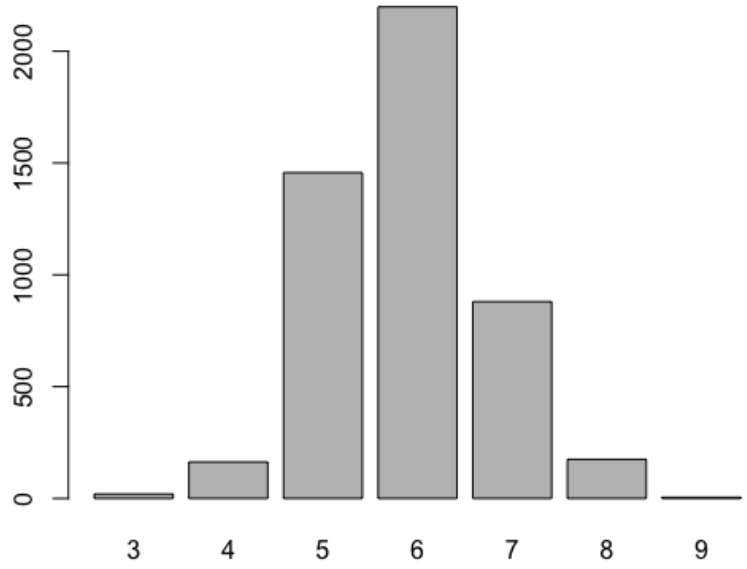

Figura 1: Distribuzione variabile "qualità"

Dopo aver convertito i valori target (1-10) in etichette di qualità, la loro distribuzione è mostrata in figura 2.

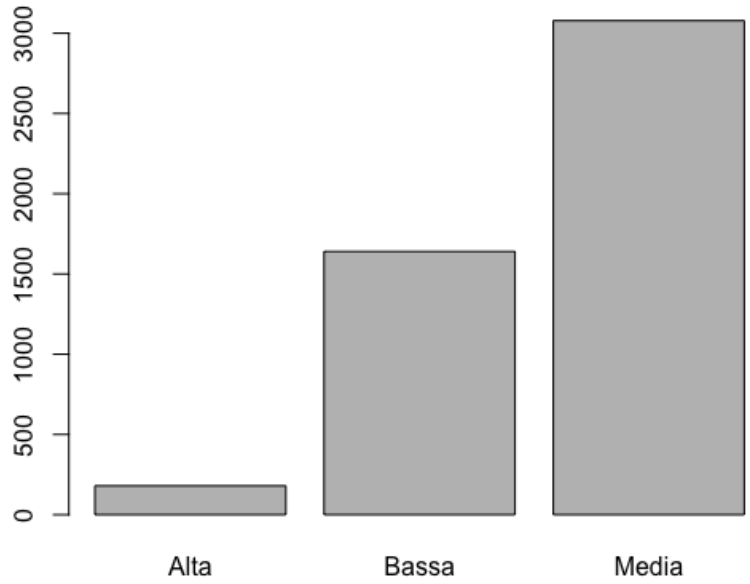

Figura 2: Distribuzione variabile "qualità" dopo conversione

2.1 Analisi univariata

Investighiamo ulteriormente producendo boxplot per ciascuna delle variabili. La figura 3 dimostra che tutte le variabili, eccetto l'*alcool*, contengono outliers.

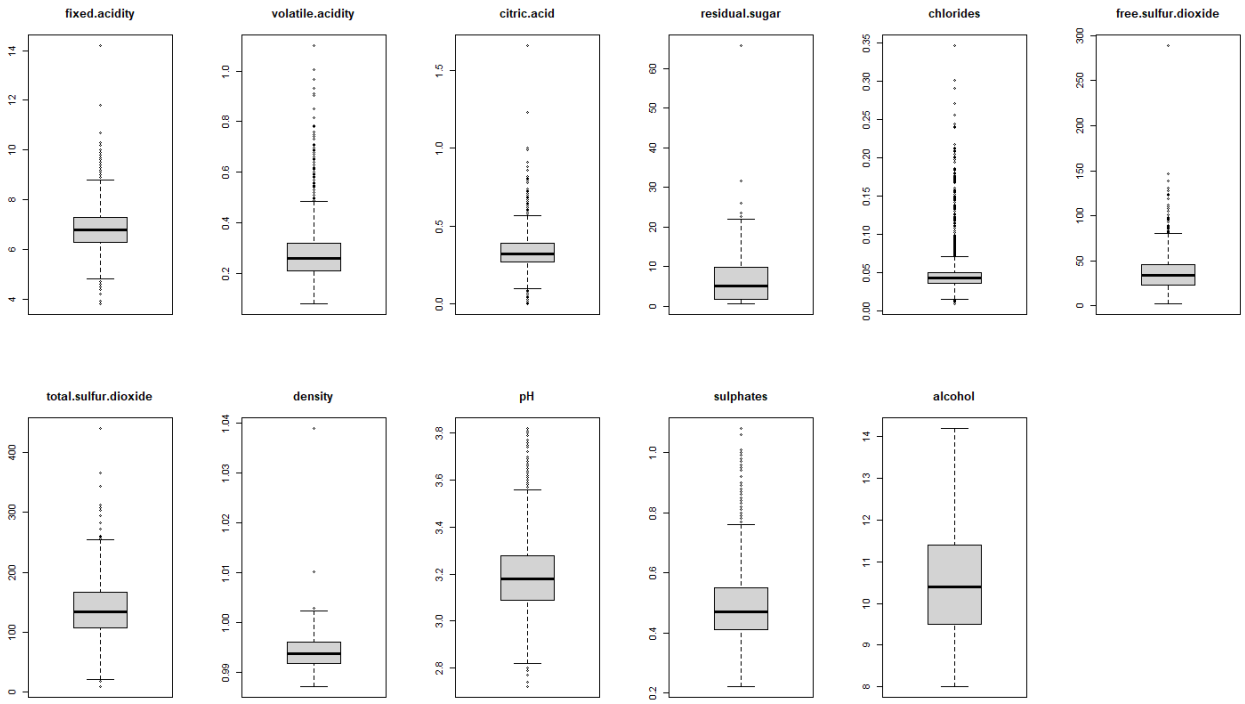

Figura 3: Outliers attributi

I seguenti grafici, in figura 4, mostra come si distribuiscono gli attribuiti in riferimento alla variabile target. Possiamo notare come la distribuzione dell'attributo *alchol* varia in base alla qualità del vino. In particolare notiamo che all'aumentare della percentuale alcolica, la qualità del vino, aumenta e al diminuire della percentuale alcolica la qualità si abbassa.

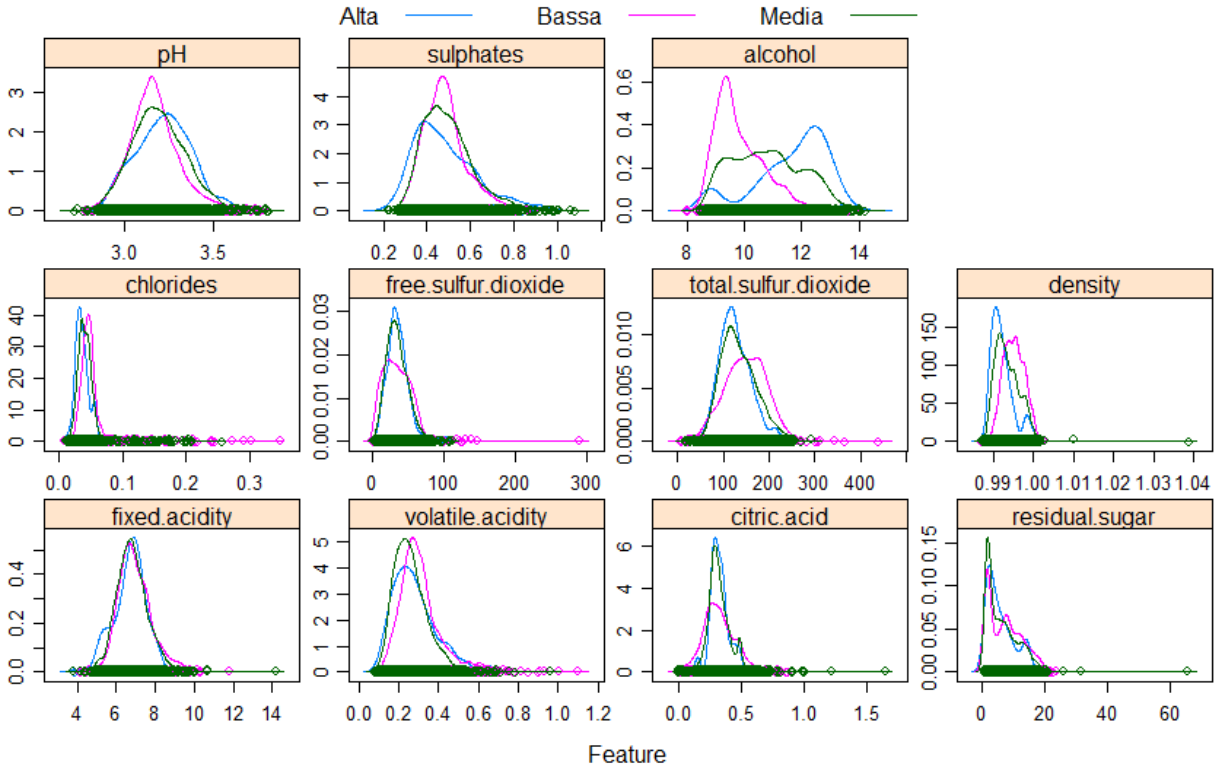

Figura 4: Distribuzione variabili in relazione alla variabile target

Osservando la distribuzione dei valori predittivi, in figura 5 e 6, notiamo che quasi tutte le distribuzioni sono positivamente distorte. *Qualità* e *pH* sono distribuiti approssimativamente normalmente.

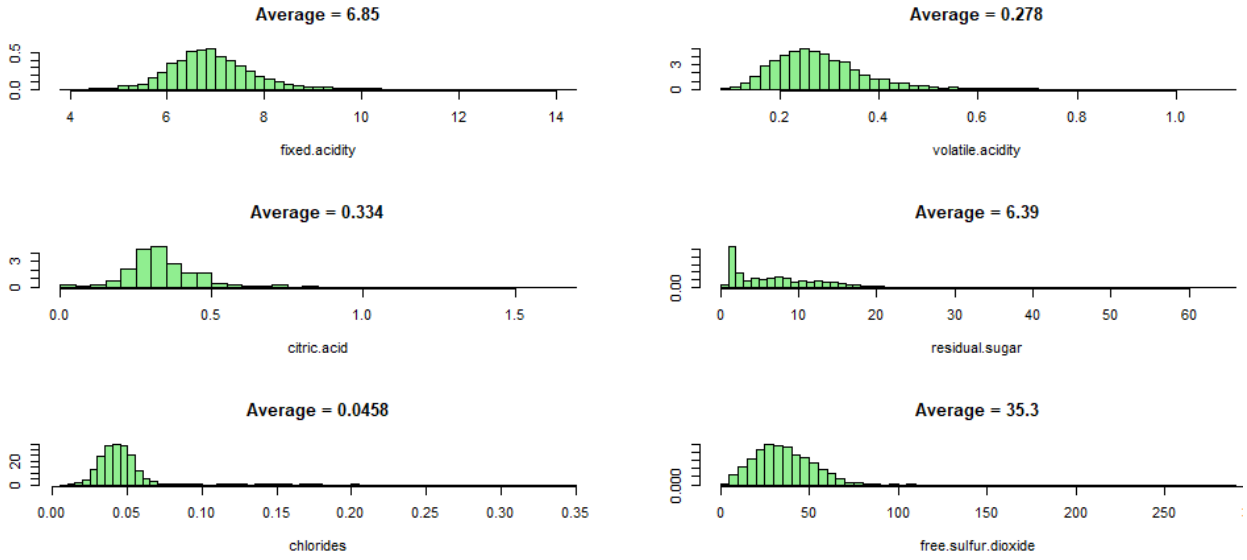

Figura 5: Distribuzione covariate

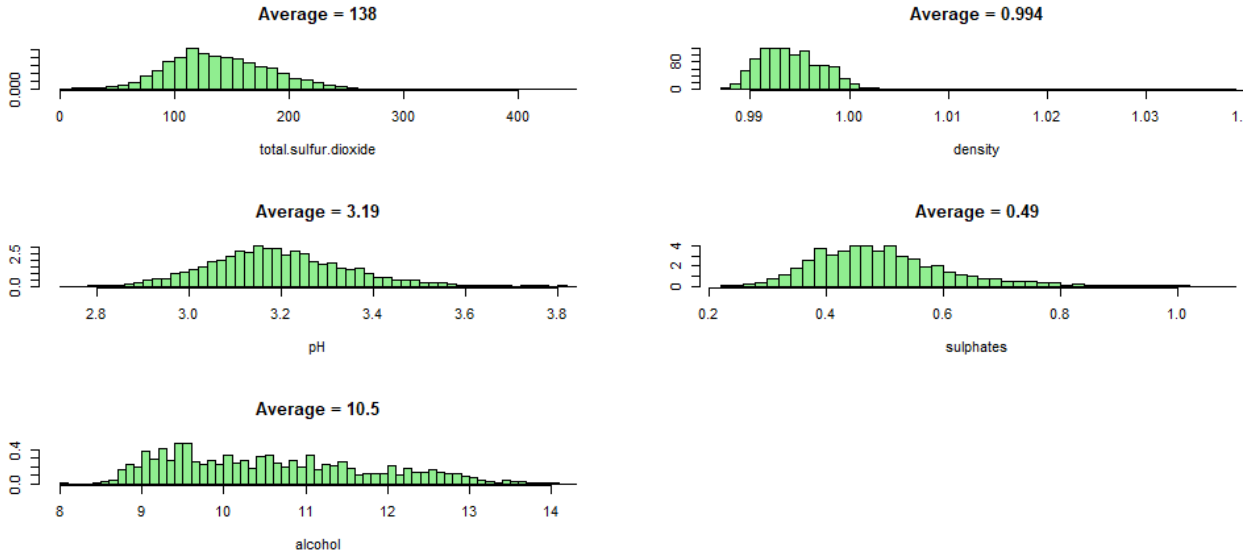

Figura 6: Distribuzione covariate

2.2 PCA

Principal Component Analysis, o PCA, è un metodo di riduzione della dimensionalità che viene spesso utilizzato per ridurre la dimensionalità di grandi insiemi di dati, trasformando un ampio insieme di variabili in uno più piccolo che contiene ancora la maggior parte delle informazioni presenti nel dataset. La riduzione del numero di variabili di un set di dati va naturalmente a scapito della precisione, ma il trucco nella riduzione della dimensionalità è scambiare un po 'di accuratezza con semplicità. Perché i set di dati più piccoli sono più facili da esplorare e visualizzare e rendono l'analisi dei dati molto più semplice e veloce per gli algoritmi di apprendimento automatico senza variabili estranee da elaborare.

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	3.22225389	29.293217	29.29322
Dim.2	1.57523993	14.320363	43.61358
Dim.3	1.22167134	11.106103	54.71968
Dim.4	1.01852235	9.259294	63.97898
Dim.5	0.97333458	8.848496	72.82747
Dim.6	0.93874151	8.534014	81.36149
Dim.7	0.72659802	6.605437	87.96692
Dim.8	0.59935848	5.448713	93.41564
Dim.9	0.41414367	3.764942	97.18058
Dim.10	0.28948714	2.631701	99.81228
Dim.11	0.02064909	0.187719	100.00000

Possiamo notare che per mantenere un livello di varianza spiegata soddisfacente (> 95%) dobbiamo mantenere 9 componenti su 11. Non avendo una notevole riduzione del numero di attributi abbiamo deciso di proseguire mantenendo tutte le componenti in modo da cercare di mantenere una maggiore accuratezza nella predizione.

2.3 Analisi multivariata

2.3.1 Matrice di Correlazione

Successivamente osserviamo le correlazioni tra le variabili con cui stiamo lavorando, attraverso una matrice di correlazione. Questo ci permette di avere una migliore comprensione delle relazioni tra le variabili grazie alla matrice di correlazione, figura 7. La *qualità* del vino è altamente correlata alla *quantità* e alla *densità* di alcol. Tuttavia, *alcol* e *densità* sono correlati negativamente. Pertanto,

uno di questi può essere utilizzato come indicatore della qualità del vino. Inoltre, è la quantità di alcol che riduce la densità, a causa della chimica, quindi la quantità di alcol è una buona scelta come indicatore della qualità del vino.

La qualità del vino è correlata negativamente con l'acidità volatile, in quanto livelli troppo alti portano a un gusto acetoso, supportando la descrizione del set di dati. La SO2 libera e la SO2 totale sono altamente correlate tra loro.

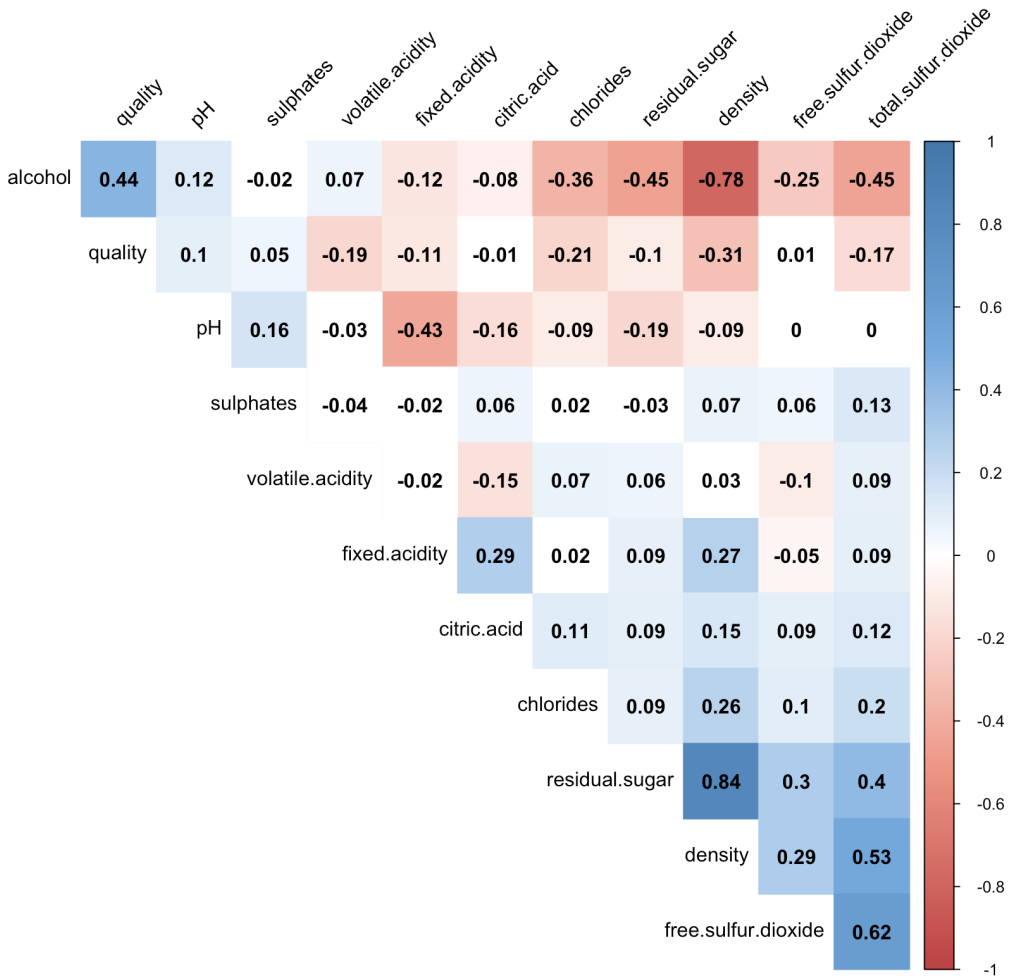

Figura 7: Matrice di correlazione

2.3.2 Relazioni tra le variabili

Le matrici di grafico a dispersione, in figura 8, sono un ottimo modo per determinare approssimativamente se si dispone di una correlazione lineare tra più variabili. Abbiamo deciso di visualizzare questa matrice tra le variabili che erano risultate maggiormente correlate (positivamente e negativamente) nella matrice di correlazione, in relazione all'obiettivo finale di classificare i vini.

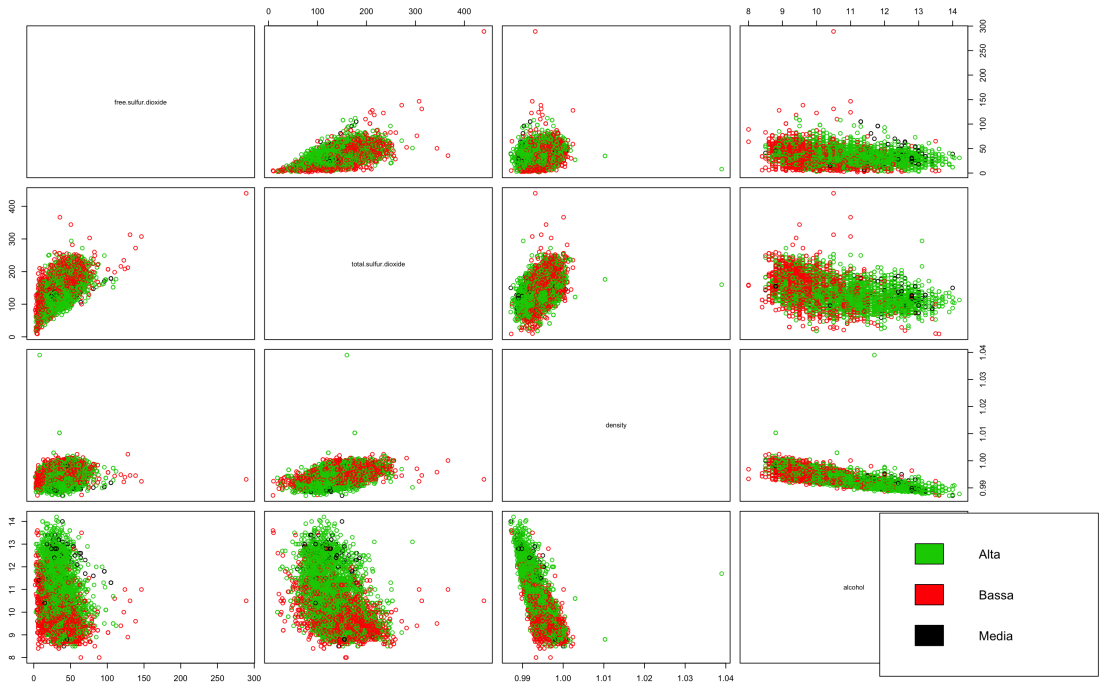

Figura 8: Grafico dispersione tra più variabili

Alchol e qualità I vini di alta qualità hanno livelli di alcol in volume più elevati, rispetto ai vini di qualità bassa e normale (fig. 9). Inoltre, alcuni vini di bassa qualità hanno quantità alcoliche elevate, rappresentati dagli outliers.

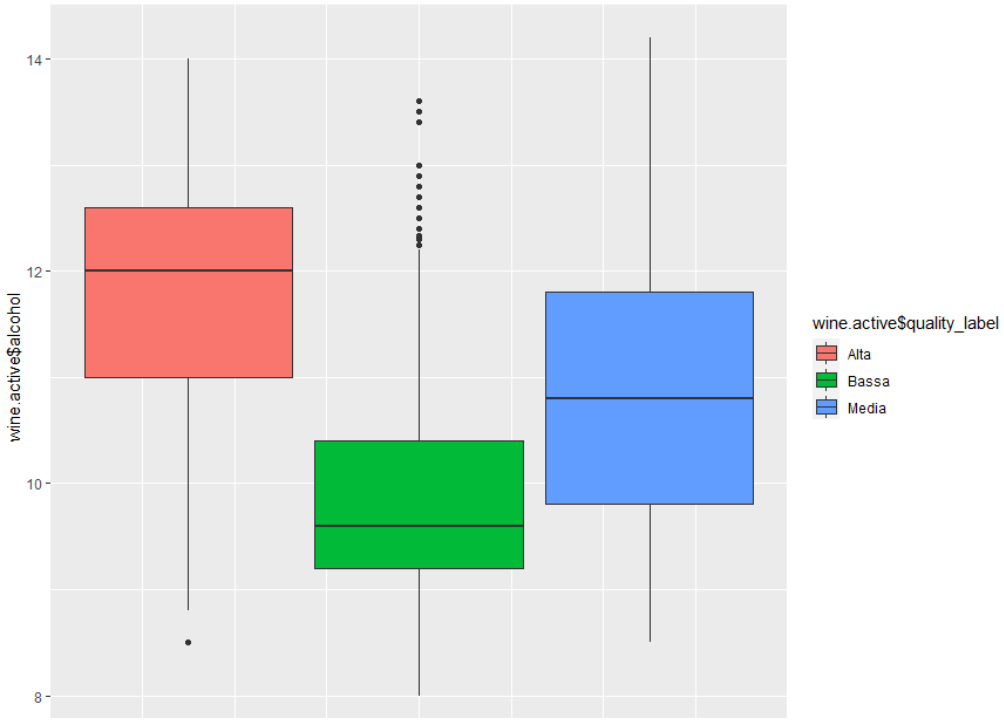

Figura 9: Distribuzione alchol in riferimento alla qualità del vino

2.3.3 Analisi bivariata

I plot mostrati in seguito cercano di evidenziare la correlazione tra l'attributo *alchol* e diversi altri attributi.

Alchol e densità La figura 10, mostra come al diminuire della percentuale alcolica e all'aumento della densità corrisponda una qualità più bassa del vino.

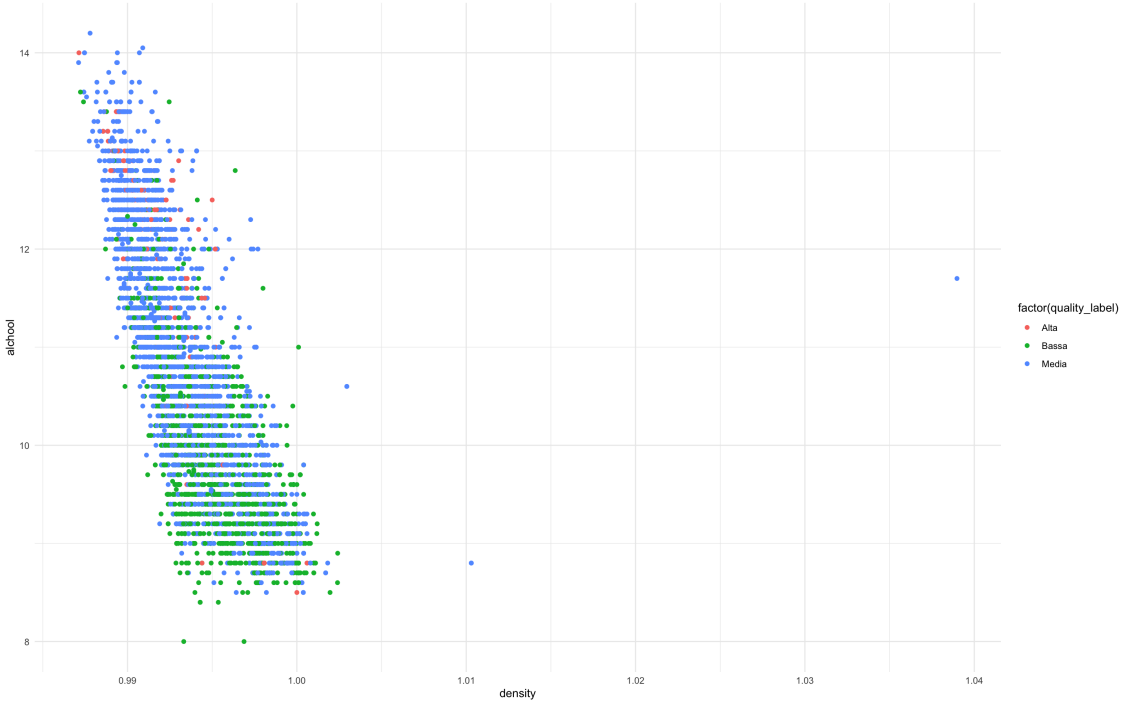

Figura 10: Relazione tra densità e alchol e la loro classificazione

Alchol e acidità volatile La figura 11, mostra come all'aumentare dell'acidità volatile la qualità del vino diminuisca. Si può notare anche in questo caso che i vini di media/alta qualità hanno tutti una percentuale alcolica maggiore e si nota che l'acidità volatile è bassa. I vini di qualità bassa hanno un'acidità volatile maggiore, Un elevato quantitativo di volatile può essere dovuto anche ad uve poco sane o da altre alterazioni microbiologiche ad opera di batteri, che avvengono quando il vino è esposto eccessivamente all'aria.

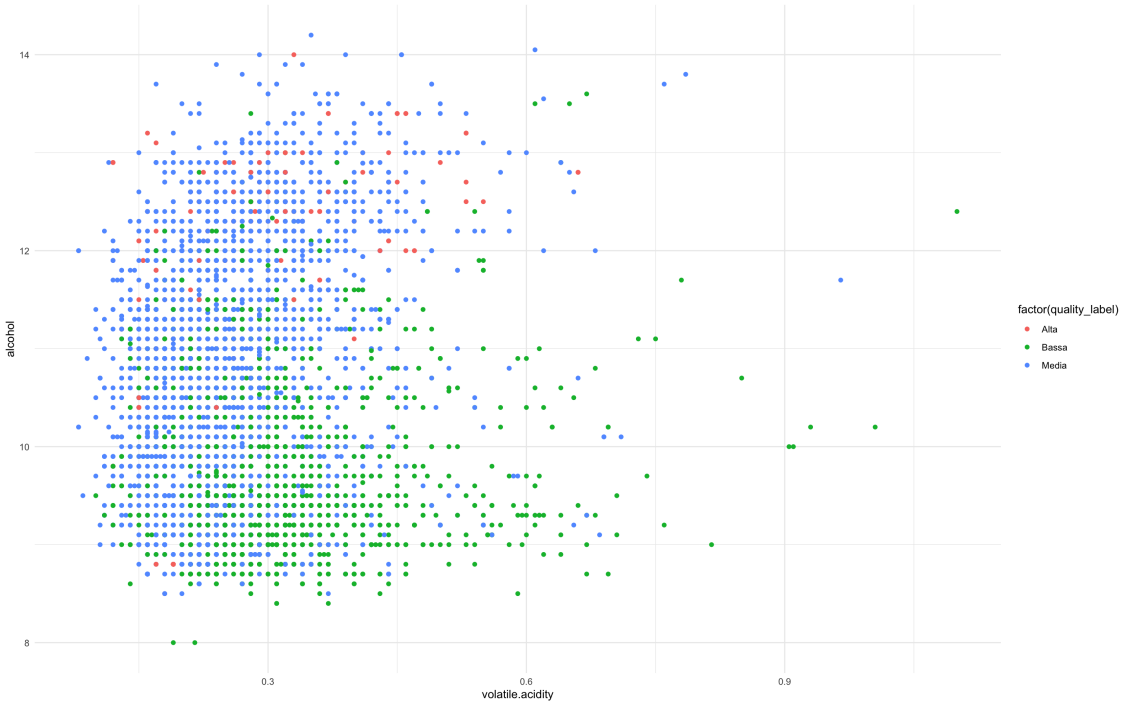

Figura 11: Relazione tra alchol e acidità volatile e la loro classificazione

3 Modelli di machine learning

Avendo sostituito al valore indicante la qualità la corrispondente etichetta (bassa, media, alta), ci troviamo di un fronte a un problema di classificazione; nello specifico un problema di apprendimento supervisionato. Abbiamo scelto di utilizzare due modelli di classificazione (Support Vector Machine e Random Forest), di cui analizzeremo i risultati. Prima di procedere effettuiamo il partizionamento del dataset, dividendolo in un sottoinsieme dedicato al training dei modelli (riservato all'incirca il 70% del volume del dataset) e la parte rimanente, ossia il 30% del volume dei dati, viene dedicato al testing dei modelli di apprendimento.
Inoltre abbiamo deciso di effettuare la classificazione su due diverse varianti del dataset, la differenza la troviamo nel criterio di assegnazione della label dunque nella variabile *quality_label*.

- Nel dataset completo la variabile *quality_label* ha i seguenti valori: Alta, Bassa, Media.
- Nel dataset ridotto abbiamo cambiato il criterio con il quale vengono assegnate le label: i vini con qualità inferiore a 7 avranno label "Bassa" mentre quelli con qualità ≥ 7 assumeranno la label "Alta".

3.1 Support Vector Machine - dataset completo

Il primo modello di apprendimento automatico tentato prevede l'utilizzo di SVM. Il kernel più comune applicato per i dati che non sono separabili linearmente è il kernel *radiale*, utilizzato quindi anche nel nostro modello. Dopo aver effettuato il training del modello usando i valori di default `cost=1`, `gamma=1` e la prediction sul testset abbiamo ottenuto il seguente risultato:

```
Confusion Matrix and Statistics

        Alta Bassa Media
  Alta    10     0    43
  Bassa    0   249   234
  Media    1    69   862

               Accuracy : 0.7636
```

Per cercare di migliorare le performance di questo modello abbiamo applicato il tuning dei parametri. L'oggetto restituito, contenente i valori ottimali, indica nella componente `best.parameters`: `cost=1, gamma=0.5`.

Dopo aver eseguito il train e il test del modello usando i nuovi parametri, questa è la matrice di confusione risultante:

```
Confusion Matrix and Statistics
        Alta Bassa Media
  Alta     3     1    49
  Bassa    0   302   181
  Media    1   110   821
               Accuracy : 0.767
```

Nel modello con i parametri ottimali si può notare un leggero aumento dell'**accuratezza** (+0,0034). In entrambi i modelli, la SVM classifica erroneamente la maggior parte dei vini di alta qualità; tende a etichettare correttamente la maggior parte dei vini di media qualità.

3.2 Random Forest - dataset completo

Il secondo modello di apprendimento scelto è il Random Forest. Random Forest viene utilizzato per migliorare le prestazioni degli alberi decisionali aggregando molti alberi decisionali per fornire un modello ottimale che non è suscettibile all'overfitting. Negli alberi decisionali, la divisione viene eseguita utilizzando un set completo di features, ma in Random Forest viene utilizzato solo un sottoinsieme casuale di features per la divisione. Poiché costruisce molti alberi con una correlazione minore tra gli alberi, ciò ridurrebbe la varianza dell'albero ottimale. Dopo aver eseguito il train e il test del modello abbiamo ottenuto il seguente risultato:

```
Confusion Matrix and Statistics
        Alta Bassa
  Alta   144   156
  Bassa   26  1121
               Accuracy : 0.8127
```

L'accuratezza del modello è 81.27%, quindi la qualità del 18.73% dei vini verrà predetta in modo errato dal modello. Il modello con Random Forest ha mostrato un'accuratezza maggiore rispetto all'uso di SVM (+0.0457).

3.3 Support Vector Machine - dataset ridotto

La matrice di correlazione della SVM mostra ancora una volta la difficoltà a riconoscere correttamente i vini di alta qualità.

```
Confusion Matrix and Statistics
        Alta Bassa
```

```
  Alta   144   156
  Bassa   26  1121
                Accuracy : 0.8742
```

Come fatto in precedenza, eseguiamo il tuning dei parametri usando il dataset ridotto. I parametri ottimali risultanti dal tuning sono: `cost=10, gamma=1`.

Dopo aver eseguito il train e il test del modello usando i nuovi parametri, questa è la matrice di confusione risultante:

```
Confusion Matrix and Statistics
        Alta Bassa
  Alta   172   128
  Bassa   49  1098
                Accuracy : 0.8777
```

Nel modello con i parametri ottimali si può notare un leggero aumento dell'**accuratezza**. La SVM classifica erroneamente circa la metà dei vini di alta qualità. La maggior parte dei vini di bassa qualità sono etichettati correttamente.

3.4 Random Forest - dataset ridotto

Dopo aver eseguito il train e il test del modello abbiamo ottenuto il seguente risultato:

```
 Confusion Matrix and Statistics
          Alta Bassa
    Alta   184    34
    Bassa  116  1113
                Accuracy : 0.8963
```

L'accuratezza del modello usando Random Forest è leggermente maggiore (+0.0186) rispetto al modello SVM con parametri ottimali.

Dopo aver effuato la predizione sia con il dataset completo che con quello ridotto abbiamo notato che l'*accuracy* aumentava nel dataset ridotto e quindi abbiamo deciso di effettuare le operazioni di misurazione delle performance sul dataset ridotto.

4 Analisi risultati ottenuti

4.1 Risultati SVM

Precision, Recall e F-Measure dei modelli di riferimento Analizziamo le differenze tra i risultati per le due classi: "bassa" e "alta".

4.1.1 Classe "Bassa"

```
Confusion Matrix and Statistics
        Alta Bassa
  Alta   172   128
  Bassa   49  1098

               Accuracy : 0.8777
                 95% CI : (0.8597, 0.8941)
    No Information Rate : 0.8473
    P-Value [Acc > NIR] : 0.0005492
                  Kappa : 0.5878

 Mcnemar s Test P-Value : 4.55e-09

              Precision : 0.9573
                 Recall : 0.8956
                     F1 : 0.9254
            Prevalence : 0.8473
        Detection Rate : 0.7588
   Detection Prevalence : 0.7927
      Balanced Accuracy : 0.8369

       Positive Class : Bassa
```

Da tali risultati otteniamo le seguenti informazioni:

- L'**accuratezza** del modello è 0.8777. Il numero dei FP (falsi positivi) è 128 vini, mentre il numero di FN (falsi negativi) è 49 (una percentuale bassa).

- La **precisione** della predizione durante il calcolo è 0.9573, è il rapporto tra il numero delle previsioni corrette di un evento (classe) sul totale delle volte che il modello lo prevede.
- La **recall** è dello 0.8956 il che ci indica che il modello ha previsto le osservazioni positive in corrispondenza con la classe "Bassa". La recall misura la sensibilità del modello. E' il rapporto tra le previsioni corrette per una classe sul totale dei casi in cui si verifica effettivamente.
- La **f-measure** risulta del 0.9254 dunque precision e recall sono uniformi. E' una misura dell'accuratezza del test. La misura tiene in considerazione precision e recall.

4.1.2 Classe "Alta"

```
Confusion Matrix and Statistics

        Alta Bassa
  Alta   172   128
  Bassa   49  1098

               Accuracy : 0.8777
                 95% CI : (0.8597, 0.8941)
    No Information Rate : 0.8473
    P-Value [Acc > NIR] : 0.0005492
                  Kappa : 0.5878

 Mcnemar s Test P-Value : 4.55e-09

              Precision : 0.5733
                 Recall : 0.7783
                     F1 : 0.6603
             Prevalence : 0.1527
         Detection Rate : 0.1189
   Detection Prevalence : 0.2073
      Balanced Accuracy : 0.8369

       Positive Class : Alta
```

Da tali risultati otteniamo le seguenti informazioni:

- L'**accuratezza** del modello è 0.8777, ottenendo la medesima matrice di confusione.
- La **precisione** della predizione durante il calcolo è 0.5733 come si può notare dalla matrice di confusione. Questo perché vengono classificati 128 vini di bassa qualità, come alta qualità.
- La **recall** è dello 0.7783 il che ci indica che il modello ha previsto le osservazioni positive in corrispondenza con la classe "Alta".
- La **f-measure** risulta del 0.6603 il che è logico in quanto la precisione e la recall non hanno valori vicini come nel caso della classe "bassa".

4.2 Risultati Random Forest

Precision, Recall e F-Measure dei modelli di riferimento Analizziamo le differenze tra i risultati per le due classi: "bassa" e "alta".

4.2.1 Classe "Bassa"

```
Confusion Matrix and Statistics

                 Alta Bassa
            Alta   184    34
            Bassa  116  1113

               Accuracy : 0.8962
                 95% CI : (0.8751, 0.9077)
    No Information Rate : 0.7927
    P-Value [Acc > NIR] : < 2.2e-16
                  Kappa : 0.6342

 Mcnemar s Test P-Value : 3.026e-11

              Precision : 0.9025
                 Recall : 0.9686
                     F1 : 0.9344
             Prevalence : 0.7927
         Detection Rate : 0.7678
```

```
   Detection Prevalence : 0.8507
      Balanced Accuracy : 0.7843

       Positive Class : Bassa
```

Le analisi sui risultati fatte per le informazioni presentate nella sezione 4.1.1 valgono anche per i risultati del modello Random Forest per la classa Bassa.

4.2.2 Classe "Alta"

```
Confusion Matrix and Statistics
                 Alta Bassa
            Alta   184    34
            Bassa  116  1113

               Accuracy : 0.8962
                 95% CI : (0.8743, 0.9071)
    No Information Rate : 0.7927
    P-Value [Acc > NIR] : < 2.2e-16
                  Kappa : 0.6333

 Mcnemar s Test P-Value : 1.717e-10

              Precision : 0.8265
                 Recall : 0.6033
                     F1 : 0.6975
             Prevalence : 0.2073
         Detection Rate : 0.1251
   Detection Prevalence : 0.1513
      Balanced Accuracy : 0.7851

       Positive Class : Alta
```

Le analisi sui risultati fatte per le informazioni presentate nella sezione 4.1.2 valgono anche per i risultati del modello Random Forest per la classa Alta.

4.3 Curva ROC

Abbiamo voluto valutare l'andamento della curva ROC riferita al modello di SVM allenato utilizzando il subset del training set. Quindi abbiamo allenato il modello calcolando le probabilità delle etichette ed eseguito tale modello sul subset del test set. Rappresentata in seguito in figura 12.

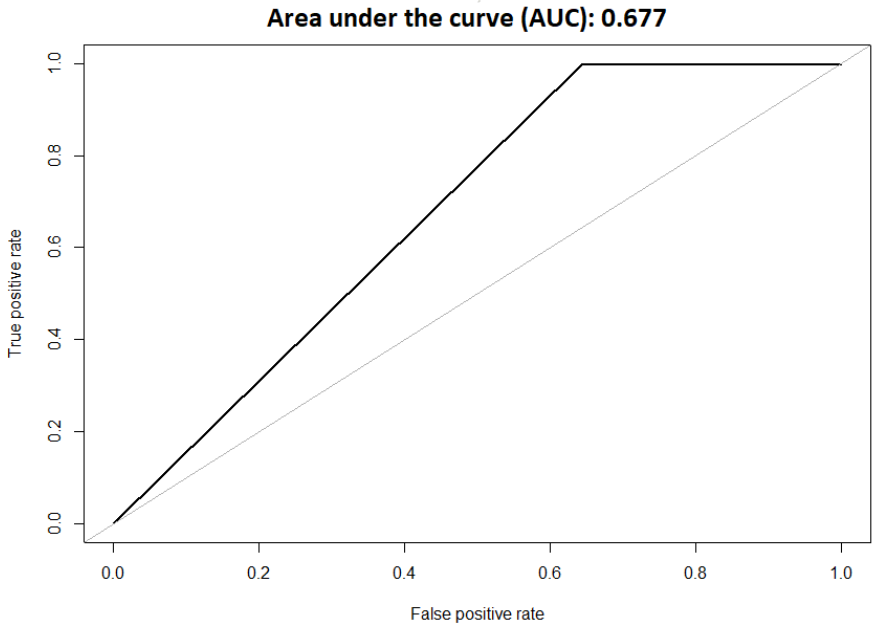

Figura 12: Curva ROC SVM

Abbiamo eseguito la stessa valutazione per il modello Random Forest, mostrato in figura 14.

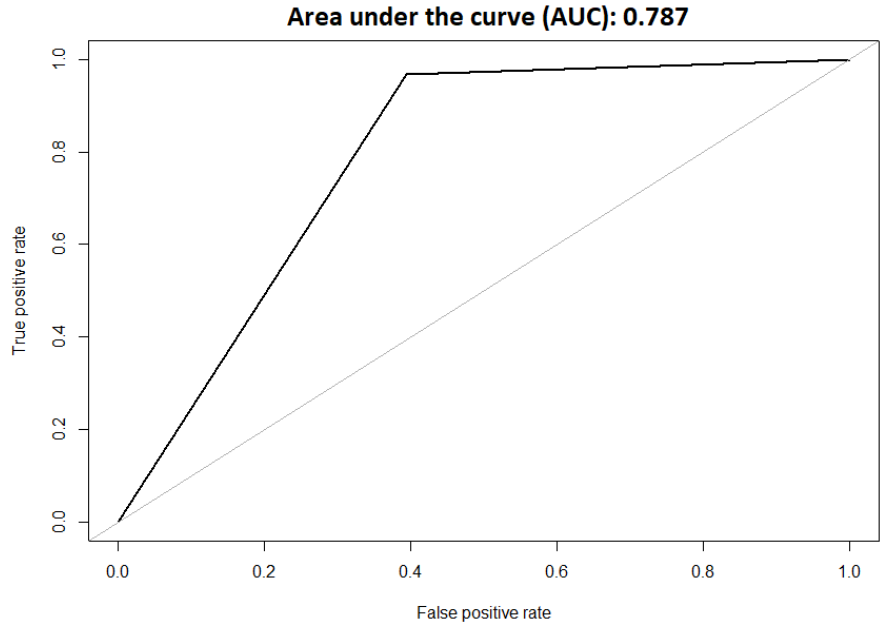

Figura 13: Curva ROC Random Forest

4.4 Confronto tra i modelli

In quest'ultima fase abbiamo messo in relazione i modelli per confrontarne le performance. Abbiamo eseguito una 10-fold cross validation con 3 ripetizioni sui seguenti modelli:

- SVM allenata sul trainset, del dataset ridotto, con parametri gamma = 1 e costo = 10.
- Random Forest su trainset del dataset ridotto.

In seguito all'allenamento dei modelli e alla predizione delle probabilità sulle etichette "Alta" e "Bassa", siamo andati a computare e visualizzare con un plot unico le ROC relative a ogni modello.

L'immagine 14 rappresenta le 2 ROC. La curva rossa rappresenta il modello SVM, la curva blu si riferisce al modello Random Forest.

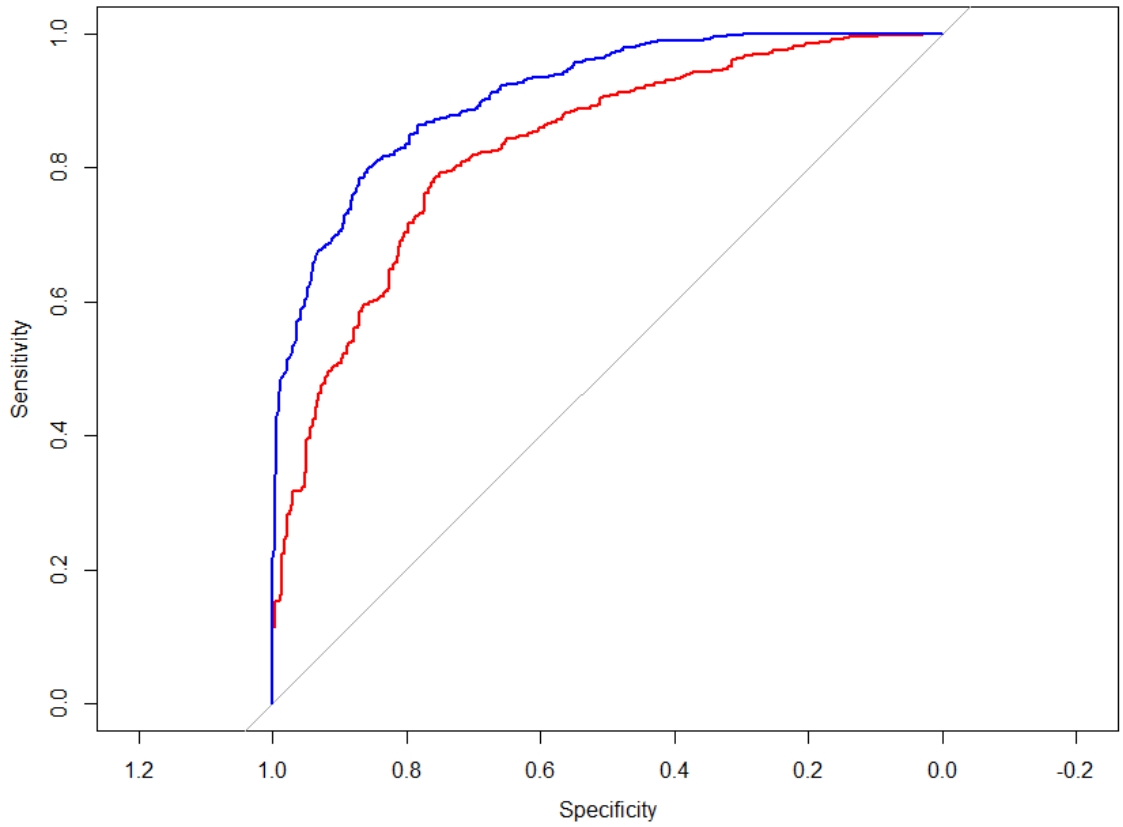

Figura 14: Roc dei due modelli

Un modello eccellente ha AUC vicino a 1, il che significa che ha una buona misura di separabilità. Un modello scadente ha AUC vicino allo 0, il che significa che ha la peggiore misura di separabilità. In effetti, significa che sta ricambiando il risultato. Prevede 0 come 1 e 1 come 0. E quando AUC è 0,5, significa che il modello non ha alcuna capacità di separazione delle classi. Quando AUC è ad esempio 0.8279, significa che c'è una probabilità del 82.79% che il modello sarà in grado di distinguere tra classe positiva e classe negativa.

I valori di AUC risultanti dall'elaborazione sono:

```
SVM             Area under the curve: 0.8279
Random Forest   Area under the curve: 0.909
```

La seguente immagine 15 mostra gli intervalli di accuratezza dei modelli, con un livello di confidenza pari al 95%.

```
Random forest   Accuracy : 0.8962
                95% CI : (0.8743, 0.9071)

SVM             Accuracy : 0.8777
                95% CI : (0.8597, 0.8941)
```

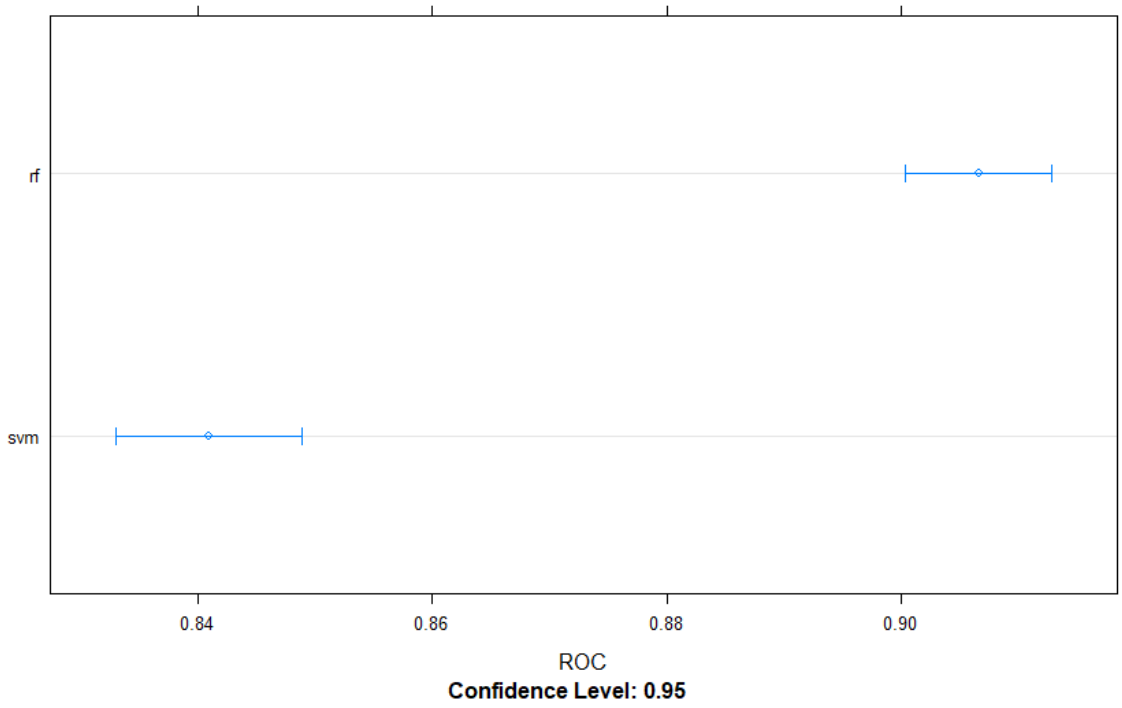

Figura 15: Intervalli di accuratezza dei modelli

L'immagine 16 mostra un confronto tra i due modelli. In particolare notiamo la stessa specificità, quindi non è un elemento discriminante di scelta tra un modello e l'altro. Possiamo notare delle differenze per quanto riguarda la ROC e la sensitività, ritenendole sufficientemente distinte.

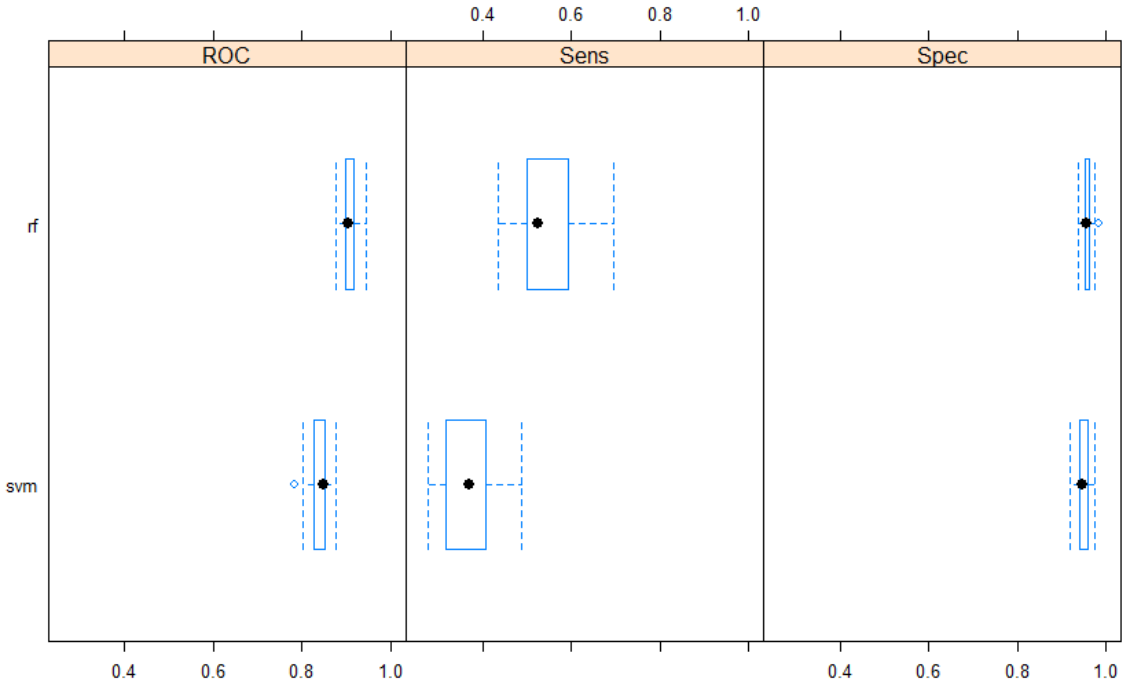

Figura 16: Plot ROC, Sensitività, Specificità dei modelli

L'ultimo confronto che abbiamo eseguito tra i modelli è quello sul tempo relativo alla computazione.

```
    Everything FinalModel Prediction
rf      185.21       1.72         NA
svm     132.47       1.58         NA
```

Da questo ultimo confronto possiamo notare che i modelli non presentano tempi computazionali molto differenti, dunque il tempo computazionale non influenza la scelta del modello da utilizzare.

5 Conclusioni

Con questo progetto abbiamo messo in pratica e compreso meglio i concetti legati all'apprendimento automatico. Ci ha permesso inoltre di incrementare le conoscenze del linguaggio di programmazione R e del suo editor RStudio.

I modelli utilizzati hanno presentato un risultato simile per quanto riguarda l'accuratezza della predizione. In particolare però abbiamo notato che, a seconda della classe positiva di riferimento, i valori di precisione variano. Infatti, scegliendo come classe positiva "Bassa", SVM ha restituito una precisione del 95.73% (90.25% per Random Forest); scegliendo come classe positiva "Alta", Random Forest ha restituito una precisione del 82.65% (contro 57.33% per SVM).

Tuttavia, questa analisi presenta alcune limitazioni. In primo luogo, il problema principale derivava dal fatto che il nostro set di dati era sbilanciato. La maggior parte dei valori di qualità erano "regolari" (5 e 6), il che non ha contribuito in modo significativo alla ricerca di un modello ottimale. Questi valori hanno reso più difficile identificare la diversa influenza di ciascun fattore su una qualità "alta" o "bassa" del vino, che era l'obiettivo principale di questa analisi.
Per migliorare il nostro modello predittivo, abbiamo bisogno di dati più equilibrati. Un altro limite che vale la pena menzionare dal set di dati era che aveva solo 12 attributi, che possono restringere la precisione della nostra previsione della qualità del vino bianco.

La soluzione per questo è includere caratteristiche dei dati più rilevanti, come l'anno di raccolta, l'ora di infusione, il luogo o il tipo di vino. In futuro, possiamo anche provare altre misurazioni delle prestazioni e altre tecniche di apprendimento automatico per migliorare le prestazioni e il confronto dei risultati.