



UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

Scuola di Scienze

Dipartimento di Informatica, Sistemistica e Comunicazione

Corso di laurea in Informatica

Progetto Machine Learning

Red Wine Quality

Magazzù Giuseppe

829612

Magazzù Gaetano

829685

Malanchini Mirco

829889

Anno Accademico 2020-2021

Indice

1	Introduzione	1
1.1	Dataset	2
2	Assunzioni e Ipotesi	4
3	Analisi Esplorativa	6
3.1	Dataset	7
3.2	Distribuzione delle variabili	8
3.3	Analisi Outlier	10
3.3.1	IQR	10
3.3.2	Winsorizing (Percentile Capping)	11
3.3.3	Grafici	15
3.4	Correlazione	26
3.5	Analisi delle componenti principali	28
4	Pre Processing	30
5	Modelli	32
5.1	CART (Classification And Regression Tree)	32
5.2	SVM (Support Vector Machine)	33
6	Esperimenti	35
6.1	Confronto tra i kernel per SVM	36
6.2	Confronto fra CART e SVM Radiale	41
6.3	Confronto fra i modelli scelti	46
7	Conclusioni	49

Capitolo 1

Introduzione

Il vino è una bevanda alcolica ottenuta dalla fermentazione del frutto della vite, dell'uva o del mosto.

Questo è un prodotto molto rinomato e conosciuto dalle diverse proprietà alimentari molto ricercate.

Queste proprietà che definiscono la qualità del vino dipendono da numerosi fattori come la variante di uva usata, il territorio in cui viene coltivata l'uva, la qualità della produzione, il tempo di fermentazione e di invecchiamento del vino.

Questi fattori influiscono fortemente anche sul costo del vino per questo motivo sono presenti numerosi studi e analisi sulle varie tipologie di vini ed uve per poter comprendere quali caratteristiche chimiche influiscono e in che modo.

Com'è intuitivo pensare, alcune proprietà sono ricercate in quanto aumentano la qualità e di conseguenza il valore del vino, come ad esempio i *polifenoli* e le *anthocianine* ricercate in quanto sostanze che migliorano il gusto e hanno un effetto positivo sulla salute.

Invece altre caratteristiche influiscono in modo negativo sul vino per quanto riguarda il gusto e possono portare ad effetti anche tossici per la salute, come ad esempio l'*anidride solforosa* che per legge deve essere al di sotto di una soglia massima perché altamente tossica per l'organismo [4].

L'analisi del vino è di centrale importanza, ma risulta molto complicata per una serie di ragioni (successivamente elencate e spiegate); inoltre si possono effettuare diverse tipologie di analisi in base al tipo di informazione cercata come ad esempio la qualità o la presenza di sostanze nocive.

Alcune delle ragioni che rendono complessa l'analisi sono:

- L'alto costo delle analisi e il tempo richiesto per effettuarle.
- Nella maggior parte delle analisi risulta troppo costoso ed elaborato considerare tutti i fattori chimici presenti, quindi è preferibile selezionare solo quelli che risultano di maggior interesse per la tipologia di vino in analisi.
- Nella maggior parte dei casi l'analisi rende non più utilizzabile il vino e questo diventa proibitivo per vini molto pregiati e costosi.
- Anche effettuando analisi approfondite e meticolose le classiche tecniche utilizzate ottengono nella maggior parte dei casi risultati parziali o poco indicativi dati i numerosi fattori chimici e organolettici.

In questo progetto è stata analizzata la qualità dei vini comprendendo le relazioni e i relativi significati delle diverse proprietà disponibili descritte all'interno del dataset scelto [6].

1.1 Dataset

Il dataset utilizzato contiene le proprietà chimiche delle varianti rosso e bianco del vino "Vinho Verde". Questo vino è un prodotto unico della regione del Minho del Portogallo.

A causa di problemi logistici e di privacy, sono disponibili solo variabili fisico-chimiche (input) e sensoriali (output), ad esempio non ci sono dati su tipi di uva, marca del vino, prezzo di vendita del vino.

Il dataset è composto da 12 attributi (11 input + 1 output) e 1599 istanze per il vino rosso e 4898 istanze per il vino bianco, inoltre le classi di qualità che descrivono le varie tipologie di vino sono ordinate e non bilanciate perché ci sono molti più vini di media qualità rispetto a quelli di alta e bassa qualità.

Gli attributi sono i seguenti:

Input

1. **Fixed acidity (tartaric acid)** [g/dm^3]: La maggior parte degli acidi coinvolti nel vino fissi o non volatili (non evaporano facilmente)

2. **Volatile acidity (acetic acid - g/dm^3)**: La quantità di acido acetico nel vino, che a livelli troppo alti può portare a un sapore sgradevole di aceto
3. **Citric acid [g/dm^3]**: Apporta una sensazione di freschezza, contribuendo all'equilibrio gustativo del vino, inoltre esalta le caratteristiche aromatiche fruttate.
4. **Residual sugar [g/dm^3]**: La quantità di zucchero rimanente dopo l'arresto della fermentazione, è raro trovare vini con meno di $1g/dm^3$ e i vini con più di $45g/dm^3$ sono considerati dolci.
5. **Chlorides (sodium chloride) [g/dm^3]**: La quantità di sali nel vino.
6. **Free sulfur dioxide [mg/dm^3]**: Una parte di anidride solforosa, detta libera, si trova sotto forma di gas o allo stato di combinazioni inorganiche (H_2SO_3 , HSO_3^- e SO_3^{2-}); solo questa parte è in grado di svolgere l'azione antisettica.
7. **Total sulfur dioxide [mg/dm^3]**: La legge fissa dei limiti per l'anidride solforosa totale, la concentrazione di SO_2 totale in un vino al momento della sua immissione sul mercato deve essere inferiore a $210\ mg/dm^3$ per i vini bianchi e a $160\ mg/dm^3$ per i vini rossi.
8. **Density [g/cm^3]**: Il rapporto tra la massa e il volume.
9. **pH**: Indica l'acidità del vino. Il vino bianco ha valori ottimali compresi tra 3.00 e 3.30 pH, mentre il pH del vino rosso è solitamente compreso fra 3.40 e 3.50 pH. Il pH ottimale prima del processo di fermentazione è compreso tra 2.9 e 4.0 pH.
10. **Sulphates (potassium sulphate) [g/dm^3]**: Sono molecole composte da ossigeno e zolfo il cui compito è quello di prevenire l'ossidazione degli alimenti, svolgono quindi una funzione antiossidante e antimicrobica; inoltre possono essere presenti naturalmente oppure aggiunti.
11. **Alcohol (% by volume)**: Percentuale di alcol presente all'interno nel vino.

Output

12. **Quality (0 - 10)**: la qualità è assegnata sulla base del giudizio di esperti

Capitolo 2

Assunzioni e Ipotesi

Dalle prime analisi effettuate sul dataset si sono riscontrati alcuni problemi rispetto alla gestione delle classi di qualità.

Come già descritto in precedenza, sono presenti ben 10 differenti classi di qualità tra vini rossi e vini bianchi e questo porta a dover gestire un problema di classificazione a 10 classi.

Dopo aver osservato tramite l'analisi del dataset le varie distribuzioni dei dati si è scelto di raggruppare le classi di qualità riconducendosi a un problema di classificazione binaria.

In questo capitolo vengono mostrati i grafici e spiegate le motivazioni che hanno portato questa scelta.

- **Classificazione a 10 classi:** ovvero quella originale rappresentata nel dataset.
- **Classificazione a 2 classi:** raggruppando le classi originali in modo tale che i vini di qualità inferiore alla qualità originale 6 compresa sono vini di bassa qualità mentre i restanti fanno parte dei vini di alta qualità.

Le due tipologie sono state confrontate per poter scegliere quale risultasse la migliore. In primo luogo sono state osservate le distribuzioni rispetto al numero di istanze [2.1].

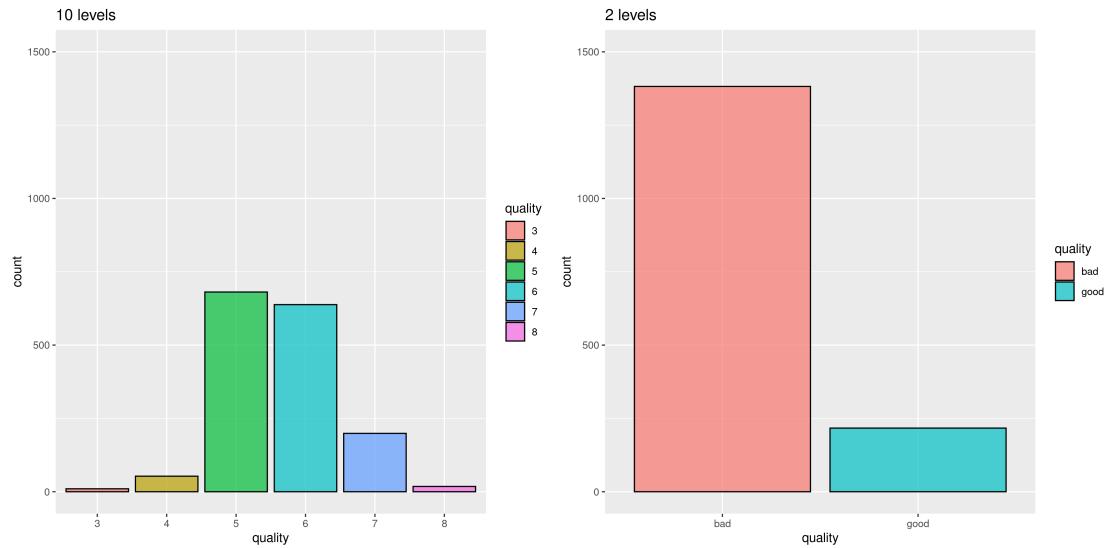


Figura 2.1: grafico che rappresenta la distribuzione dei dati rispetto alle due tipologie di classificazione prese in considerazione

Da questa analisi si è notato come le istanze non siano distribuite in modo uniforme, anzi si ha una prevalenza di dati rispetto alle qualità centrali e una scarsa rappresentazione delle qualità più basse e più alte, per questo motivo si è scartata l'ipotesi di poter sfruttare una classificazione a 10 classi, perché non in grado di rappresentare in modo appropriato le 10 classi.

Osservando il grafico [2.1] si può notare come per alcune classi di qualità non siano presenti istanze che li rappresentino.

La classificazione a due classi è stata scelta perché anche se sbilanciata ha un buon numero di istanze che rappresentano sia la qualità bassa sia la qualità alta.

Inoltre si è scelto di operare solamente usando una tipologia di vino perché operare considerando contemporaneamente vini bianchi e vini rossi rende più complessa la distinzione tra vini di bassa e alta qualità, questo per via delle loro diverse caratteristiche fisico-chimiche che li contraddistinguono.

Quindi è stata considerata solamente la porzione di dataset relativa ai vini rossi, analogamente lo stesso procedimento può essere implementato per i vini bianchi.

Capitolo 3

Analisi Esplorativa

L'Exploratory Data Analysis, spesso abbreviato in EDA, è una tecnica usata nel campo della Data Science per approfondire la conoscenza del dataset su cui si intende lavorare, operazione cruciale per svolgere su esso qualsiasi tipo di attività. Questa analisi permette di approfondire e conoscere il dataset attraverso diverse tecniche che tendono a variare per ogni dataset.

Come verrà presentato in questo capitolo i principali strumenti attraverso i quali si studia il dataset sono i grafici e i calcoli numerici di particolari valori indicativi di andamenti e correlazioni tra i dati.

I grafici che si utilizzano sono di diverse tipologie in base a ciò che si vuole catturare e percepire dal dataset.

Questo studio sui dati permette di verificare alcune supposizioni, ipotesi fatte a priori e permette di trovare informazioni difficili da notare in modo intuitivo o tramite semplici osservazioni.

Inoltre questa analisi se compiuta in modo adeguato può portare ad una conoscenza approfondita del dataset trovando tutte le principali correlazioni e le principali relazioni che intercorrono tra i dati.

3.1 Dataset

Per prima cosa è stato fatto un controllo dei valori mancanti e sono state calcolate delle statistiche descrittive per ogni variabile sul training set, in modo da avere una prima impressione dei dati. I valori calcolati sono stati riportati nella seguente tabella.

	missing	mean	sd	median	min	max	skew	kurtosis
fixed.acidity	0	6.86	0.85	6.80	3.80	14.20	0.68	2.44
volatile.acidity	0	0.28	0.10	0.26	0.08	1.10	1.67	5.73
citric.acid	0	0.33	0.12	0.32	0.00	1.66	1.35	6.80
residual.sugar	0	6.48	5.14	5.20	0.60	65.80	1.13	4.14
chlorides	0	0.05	0.02	0.04	0.01	0.35	5.12	39.50
free.sulfur.dioxide	0	35.35	17.17	34.00	2.00	289.00	1.57	13.58
total.sulfur.dioxide	0	138.36	42.60	134.00	9.00	440.00	0.42	0.78
density	0	0.99	0.00	0.99	0.99	1.04	1.11	11.52
pH	0	3.19	0.15	3.18	2.72	3.82	0.46	0.57
sulphates	0	0.49	0.11	0.47	0.22	1.06	1.01	1.64
alcohol	0	10.51	1.23	10.40	8.00	14.20	0.51	-0.67

- Non risultano esserci valori mancanti in nessuna variabile.
- Nessuna variabile ha valori negativi.
- Le variabili *free.sulfur.dioxide*, *total.sulfur.dioxide* e *residual.sugar* hanno un valore massimo molto alto rispetto alla media.

skew (asimmetria) è una misura quantificabile di quanto sia distorto un campione di dati dalla distribuzione normale, più è alto in valore assoluto il valore più i dati sono lontani dalla distribuzione normale. Inoltre il segno della *skew* indica se negativo che la media della distribuzione dei dati è spostata a sinistra rispetto alla normale, se positiva è spostata a destra [8].

kurtosis (curtosi) è la misura di una funzione "tailedness", spesso descritta visivamente dalla nitidezza dei valori di picco, la curtosi è spesso spiegata in termini di picco centrale e valori più alti di esso indicano un picco più alto e più nitido (una forma a campana più stretta) mentre valori inferiori indicano un picco inferiore e meno distinto. Se ho un valore maggiore di zero ho una curva più "appuntita" di

una normale, mentre se ho un valore minore di zero ho una curva più "appiattita" di una normale [7].

3.2 Distribuzione delle variabili

In questo capitolo si è analizzata la distribuzione per ogni singola variabile [3.1] e la distribuzione della singola variabile dividendo i valori assunti in base alle due classi [3.2], ovvero vino di bassa qualità e vino di alta qualità.

Per ogni grafico sull'asse delle ordinate si trova il range di valori assunti dalle istanze nel dataset mentre sull'asse delle ascisse si trova la stima di densità di probabilità.

La densità di probabilità si può vedere come quanta possibilità ho di avere un determinato valore considerando la classe e/o la variabile; inoltre la rappresentazione di questo valore astrae dalla numerosità di un determinato tipo di istanze.

Questo tipo di grafico può avere problemi con valori non continui, ma tenderà a mantenere una curva morbida anche con valori discreti e con valori mancanti.

Questa è un'analisi univariata, ovvero viene considerata una singola variabile alla volta, osserveremo solo una variabile per ogni grafico presentato in questo capitolo.

Non verranno prese in considerazione le relazioni tra diverse variabili, ma si cercherà di descrivere aspetti della singola variabile anche rispetto alle classi di qualità.

Il grafico delle distribuzioni permette di capire i valori che i dati tendono ad assumere, si può notare se assumono valori secondo una distribuzione standard oppure se tendono ad assumere maggiormente valori in alcuni specifici range, si può anche capire se sono presenti valori anomali.

Considerando anche le classi è possibile notare quanto i dati delle due classi sono correlati e la differenza tra le due distribuzioni.

Se una variabile tende ad avere due distribuzioni molto differenti per forma o per valori assunti allora si può pensare che la variabile rappresentata dal determinato grafico possa essere utile per distinguere le due classi di qualità.

Questo aspetto è particolarmente utile nelle fasi successive, infatti può influire sull'analisi delle componenti principali, sul modello di classificazione e anche sull'analisi dei risultati ottenuti dai modelli.

Analisi Esplorativa

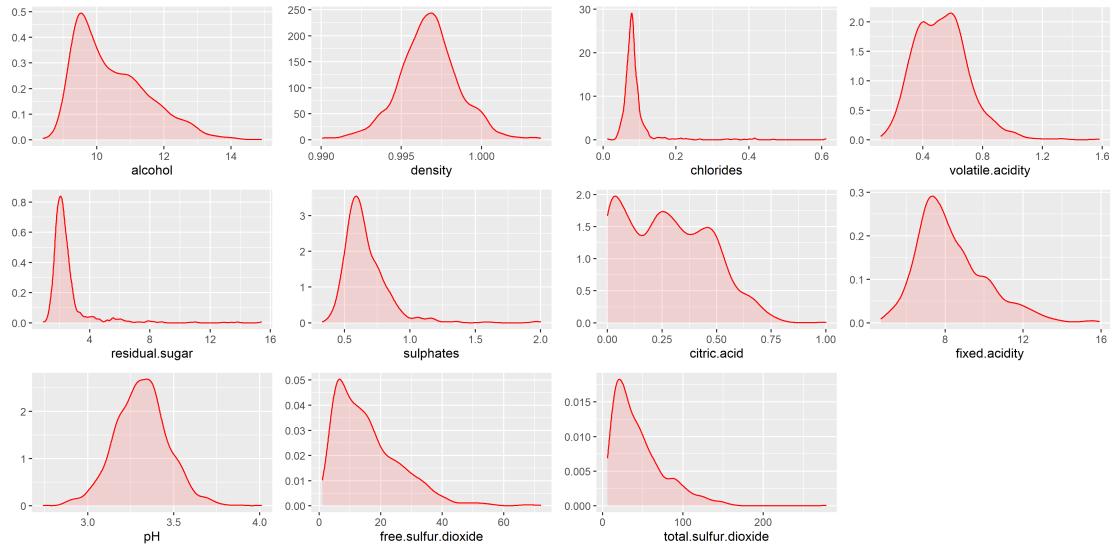


Figura 3.1: Questa immagine consiste in un insieme di grafici, dove ogni singolo grafico rappresenta la distribuzione dei valori assunti da una specifica variabile.

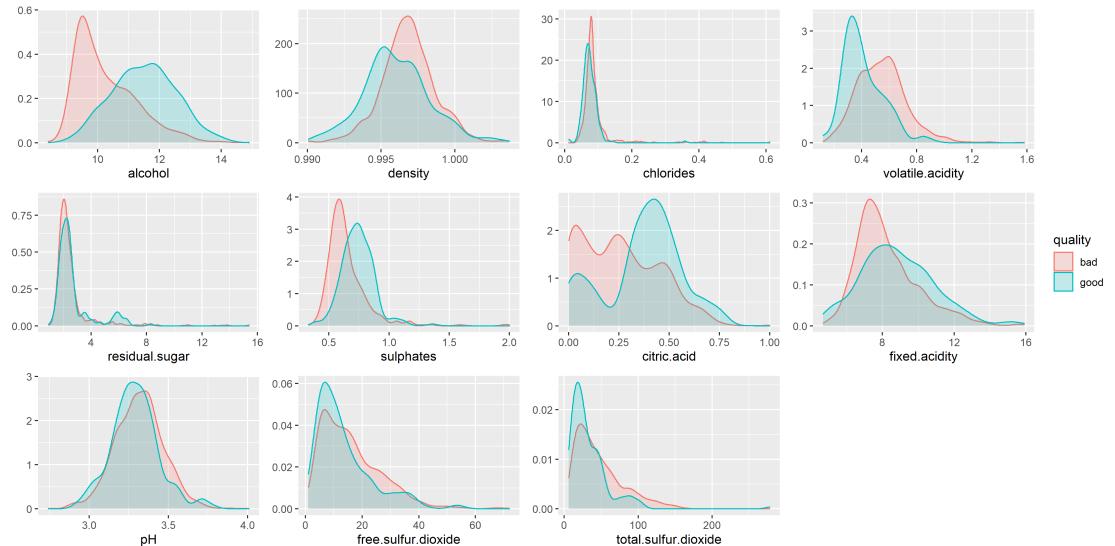


Figura 3.2: Questa immagine consiste in un insieme di grafici, dove ogni singolo grafico rappresenta la distribuzione dei valori assunti da una specifica variabile mettendo in evidenza la classe (*bad* e *good*) a cui appartengono.

Dai grafici [3.1] e [3.2] si può notare come soltanto la variabile *alcohol* abbia delle sostanziali differenze tra le due distribuzioni delle due classi e questo la rende molto interessante.

Le altre variabili tendono a non caratterizzare la differenza tra le due classi se non in minima parte. Questo indica la complessità presente nel distinguere le due classi per un possibile modello.

Questa scarsa caratterizzazione rispecchia le difficoltà, già descritte nell'introduzione [1], che si trovano nel produrre e nello svolgere le analisi.

3.3 Analisi Outlier

Gli outlier sono dei valori anomali o estremi, lontani dai valori centrali di un insieme di dati. Questi valori influenzano negativamente la media e la deviazione standard del dataset e quindi possono portare a stravolgere i risultati. Molti algoritmi di machine learning non funzionano in modo ottimale in presenza di outlier e quindi c'è bisogno di rilevarli e rimuoverli.

È stata effettuata una ricerca degli outlier su ogni attributo numerico attraverso i seguenti metodi statistici:

3.3.1 IQR

Gli outlier sono stati individuati usando l'approccio basato sul Interquartile Range (IQR). Lo scarto interquartile è un indice di dispersione, ovvero una misura di quanto i valori si allontanino da un valore centrale. Viene calcolato dalla differenza tra il terzo quartile (Q3) e il primo quartile (Q1). In questo approccio tutti i punti che si trovano al di sopra del valore $Q3 + 1.5 * IQR$ o al di sotto del valore $Q1 - 1.5 * IQR$ sono considerati outlier. Gli outlier possono essere rimossi o sostituiti con un valore fissato come ad esempio media, moda, mediana.

$$IQR = Q3 - Q1$$

$$LowerBound = Q1 - 1.5 * IQR$$

$$UpperBound = Q3 + 1.5 * IQR$$

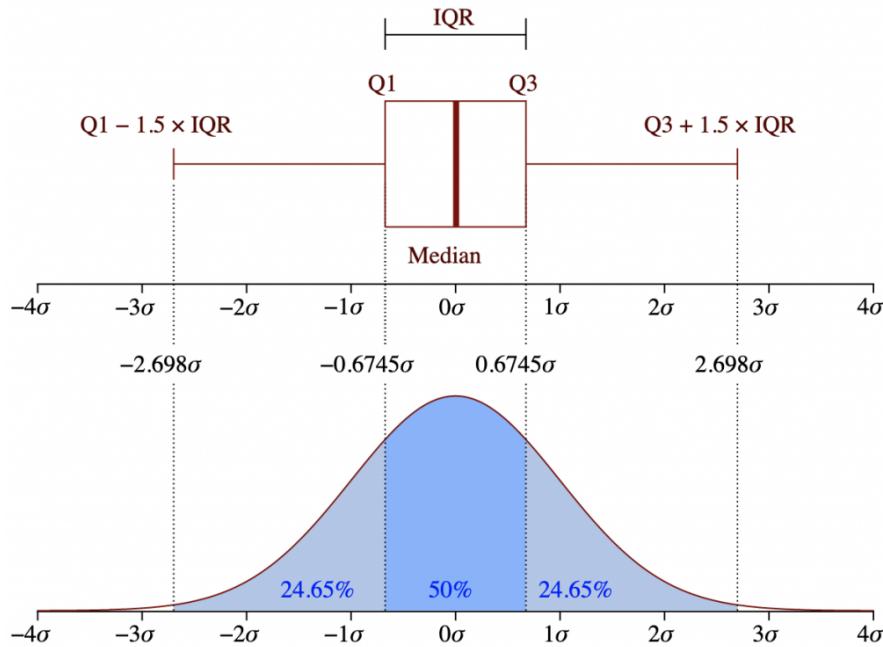


Figura 3.3: Esempio di scarto interquartile in una distribuzione normale [3]

3.3.2 Winsorizing (Percentile Capping)

Il Winsorizing è un metodo simile al metodo IQR, in questo caso si utilizzano due percentili. Tutti i valori sotto al minimo valore dell'intervallo vengono sostituiti con il minimo, e tutti i valori sopra il massimo valore dell'intervallo vengono sostituiti con il massimo. In questo lavoro sono stati usati due intervalli (5° percentile, 95° percentile) e (1° percentile, 99° percentile).

I due intervalli sono stati denominati Winsorizing 90% e Winsorizing 98%:

- Winsorizing 90% indica che il 5% inferiore dei dati viene sostituito con il 5° percentile e il 5% superiore dei dati viene sostituito con il 95° percentile.
- Winsorizing 98% indica che l'1% inferiore dei dati viene sostituito con il 1° percentile e l'1% superiore dei dati viene sostituito con il 99° percentile.

Metodo Scelto

Per decidere il metodo da usare sono stati utilizzati i boxplot. Sono state confrontate le variabili con i valori assunti dopo l'applicazione dei metodi di rimozione degli outlier (IQR, Winsorizing 90% e Winsorizing 98%). Sopra a ogni boxplot sono stati riportati i valori divisi per qualità (good, bad) per visualizzare le quantità di outlier per classe.

Nonostante il dataset sia sbilanciato, si è deciso di rimuovere completamente gli outlier poiché il numero di outliers risulta molto piccolo (circa il 3% del training set [3.4]).

In seguito sono state confrontate le distribuzioni delle variabili per ogni metodo applicato. Il metodo Winsorizing rileva un intervallo di outlier più piccolo e variabile rispetto all'IQR. Inoltre nei casi di distribuzione con distorsione laterale accumula troppi valori agli estremi, alterando così la distribuzione. Con il metodo Winsorizing 98% si risulta avere una distribuzione più smussata agli estremi. Il metodo IQR non altera la distribuzione, e rimuove un numero non troppo elevato di outliers, quindi si è deciso di usare questo metodo. Per avere un'ulteriore conferma sono stati usati dei Q-Q plot.

I Q-Q (quantile-quantile) plot sono dei grafici utili per capire se due insiemi di dati hanno la stessa distribuzione. Vengono rappresentati i punti in un piano cartesiano attraverso una coppia di quantili. Inoltre viene tracciata una retta a 45° in modo da evidenziare i punti più vicini alla retta. Due insiemi di dati hanno una distribuzione simile se i punti cadono approssimativamente sulla linea di riferimento. Analizzando i grafici si è visto che il metodo IQR ha valori più vicini alla retta, quindi si è scelto di utilizzare questo.

Nelle seguenti tabelle sono stati riportati le varie statistiche descrittive delle variabili prima e dopo la rimozione degli outlier con il metodo scelto, nelle tabelle sottostanti notiamo come rimuovendo gli outliers otteniamo un miglioramento di skew, kurtosis e della deviazione standard, mentre le altre statistiche restano praticamente invariate.

	vars	mean	sd	median	min	max	skew	kurtosis
fixed.acidity	1	8.34	1.78	7.90	4.70	15.90	0.98	1.13
volatile.acidity	2	0.53	0.18	0.52	0.12	1.58	0.71	1.46
citric.acid	3	0.27	0.19	0.26	0.00	1.00	0.32	-0.79
residual.sugar	4	2.53	1.40	2.20	0.90	15.40	4.47	27.53
chlorides	5	0.09	0.05	0.08	0.01	0.61	5.89	45.36
free.sulfur.dioxide	6	15.79	10.58	13.00	1.00	72.00	1.29	2.19
total.sulfur.dioxide	7	45.23	31.87	37.00	6.00	278.00	1.41	2.87
density	8	1.00	0.00	1.00	0.99	1.00	0.05	0.90
pH	9	3.31	0.15	3.31	2.74	4.01	0.11	0.62
sulphates	10	0.66	0.17	0.62	0.33	2.00	2.57	13.02
alcohol	11	10.45	1.07	10.20	8.40	14.90	0.84	0.13

Tabella 3.1: Prima della rimozione

	vars	mean	sd	median	min	max	skew	kurtosis
fixed.acidity	1	8.21	1.59	7.90	4.70	12.60	0.63	-0.08
volatile.acidity	2	0.52	0.17	0.51	0.12	1.00	0.27	-0.30
citric.acid	3	0.27	0.19	0.26	0.00	0.79	0.29	-0.89
residual.sugar	4	2.17	0.45	2.10	0.90	3.60	0.56	0.42
chlorides	5	0.08	0.01	0.08	0.04	0.12	0.19	0.12
free.sulfur.dioxide	6	15.11	9.33	13.00	1.00	42.00	0.79	-0.21
total.sulfur.dioxide	7	41.46	26.04	35.00	6.00	116.00	0.88	-0.06
density	8	1.00	0.00	1.00	0.99	1.00	0.01	-0.10
pH	9	3.31	0.15	3.31	2.92	3.71	0.02	-0.11
sulphates	10	0.64	0.12	0.62	0.33	0.99	0.55	-0.10
alcohol	11	10.42	1.03	10.20	8.40	13.50	0.72	-0.33

Tabella 3.2: Dopo la rimozione con IQR

Nel seguente grafico è stato confrontato il numero di outliers rimossi per variabili divisi per classe, attraverso il metodo IQR. La classe minoritaria (good) ha meno outliers come ci si può aspettare. Le variabili con più outliers rilevati sono *residual.sugar*, *chlorides*, *sulfur.dioxide* e *sulphates*. La variabile *citric.acid* ha quasi zero outliers per entrambe le classi. In totale il numero di outliers rilevati e rimossi con il metodo IQR costituiscono il 2.88% del training set.

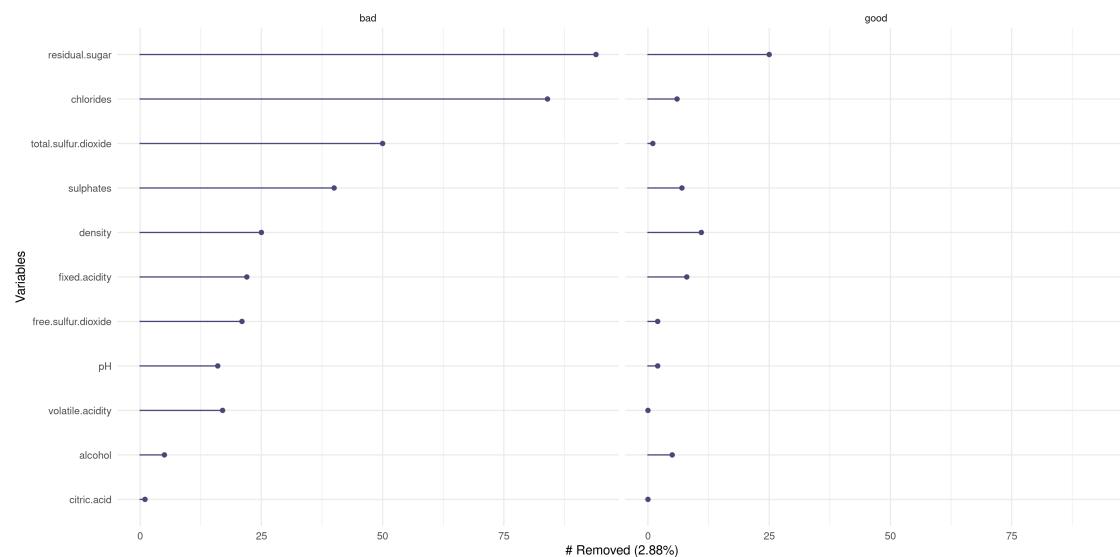
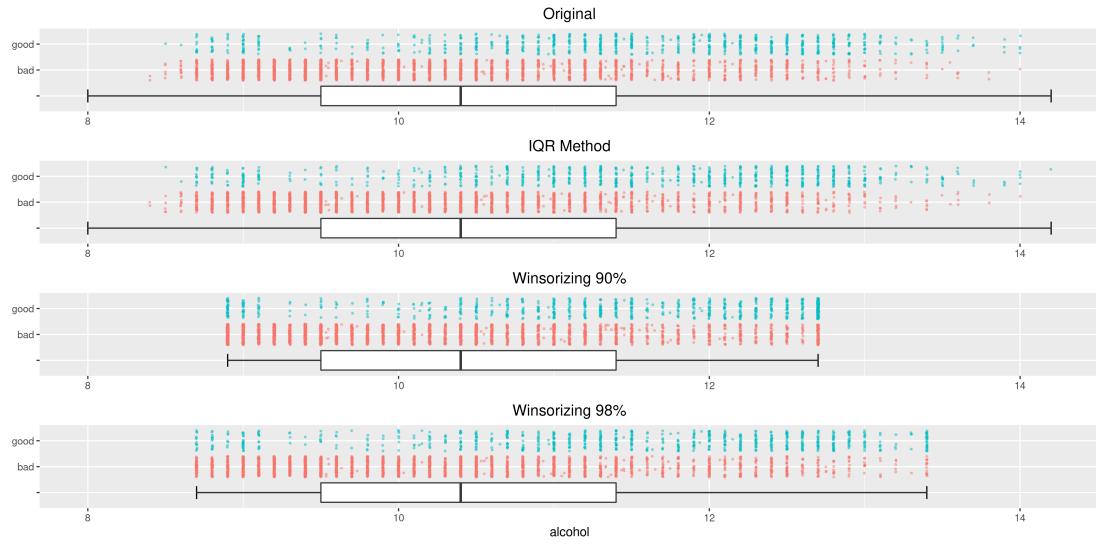


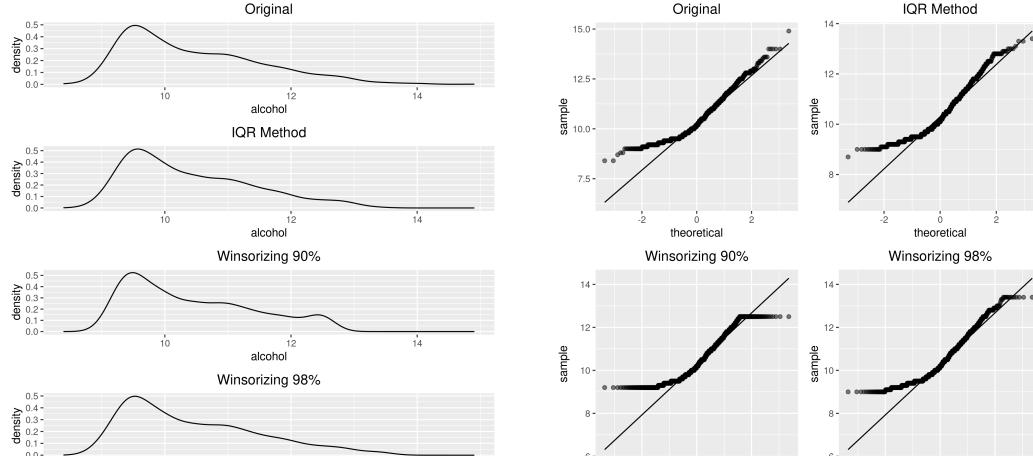
Figura 3.4: Numero di outliers rimossi per ogni variabile divisi per classe

3.3.3 Grafici

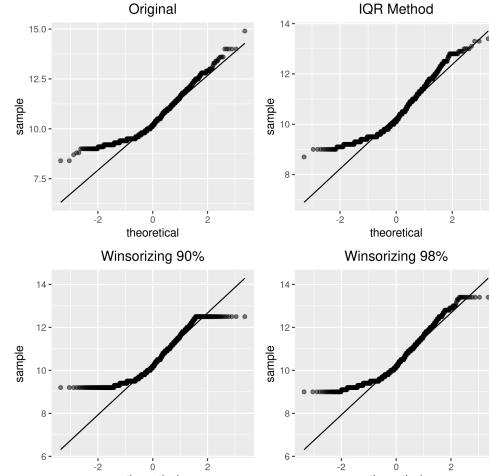
In questa sezione vengono riportati i grafici che sono stati utilizzati per l'analisi univariata degli outliers.



(a)



(b)



(c)

Figura 3.5: Alcohol

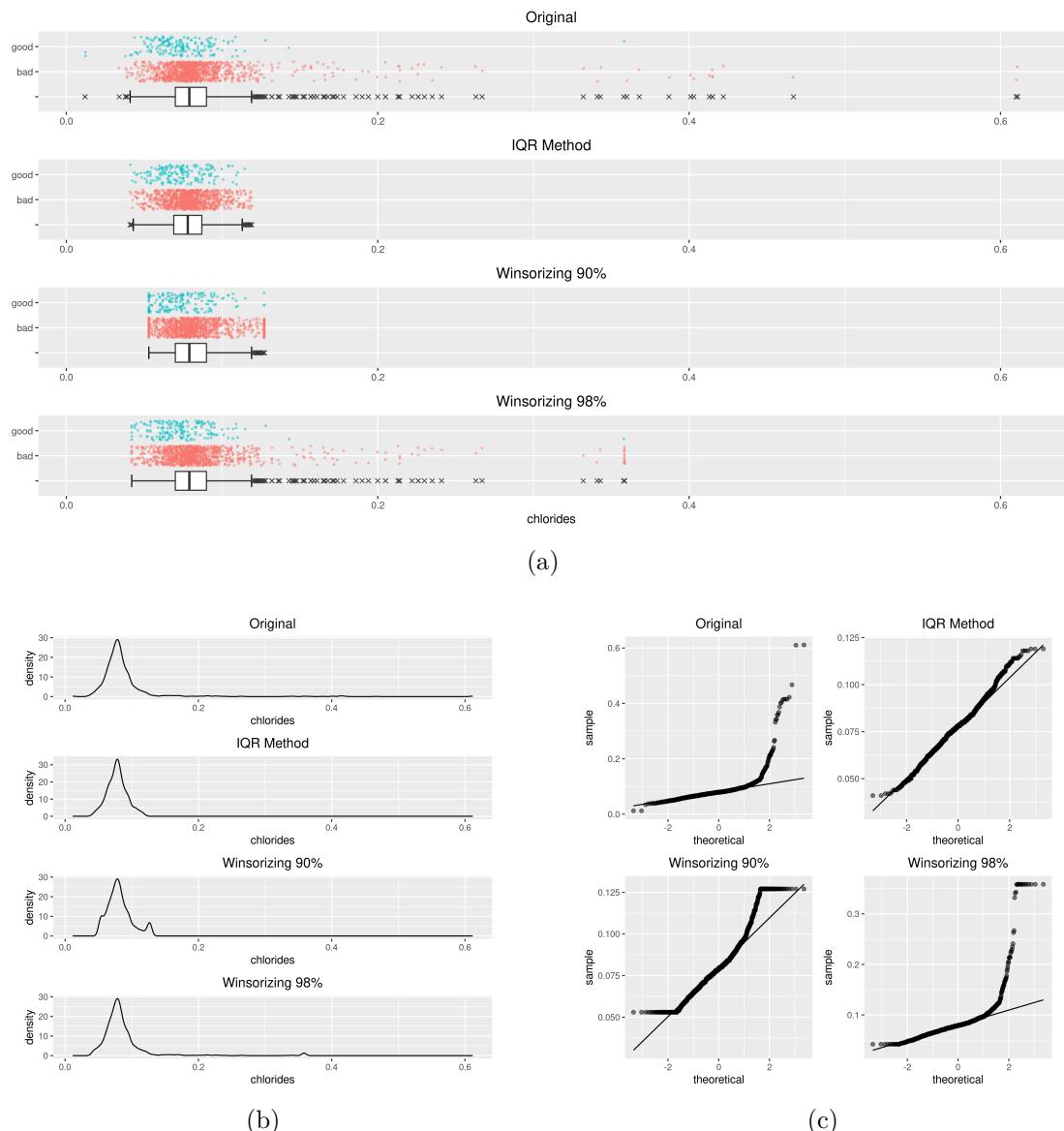


Figura 3.6: Chlorides

Analisi Esplorativa

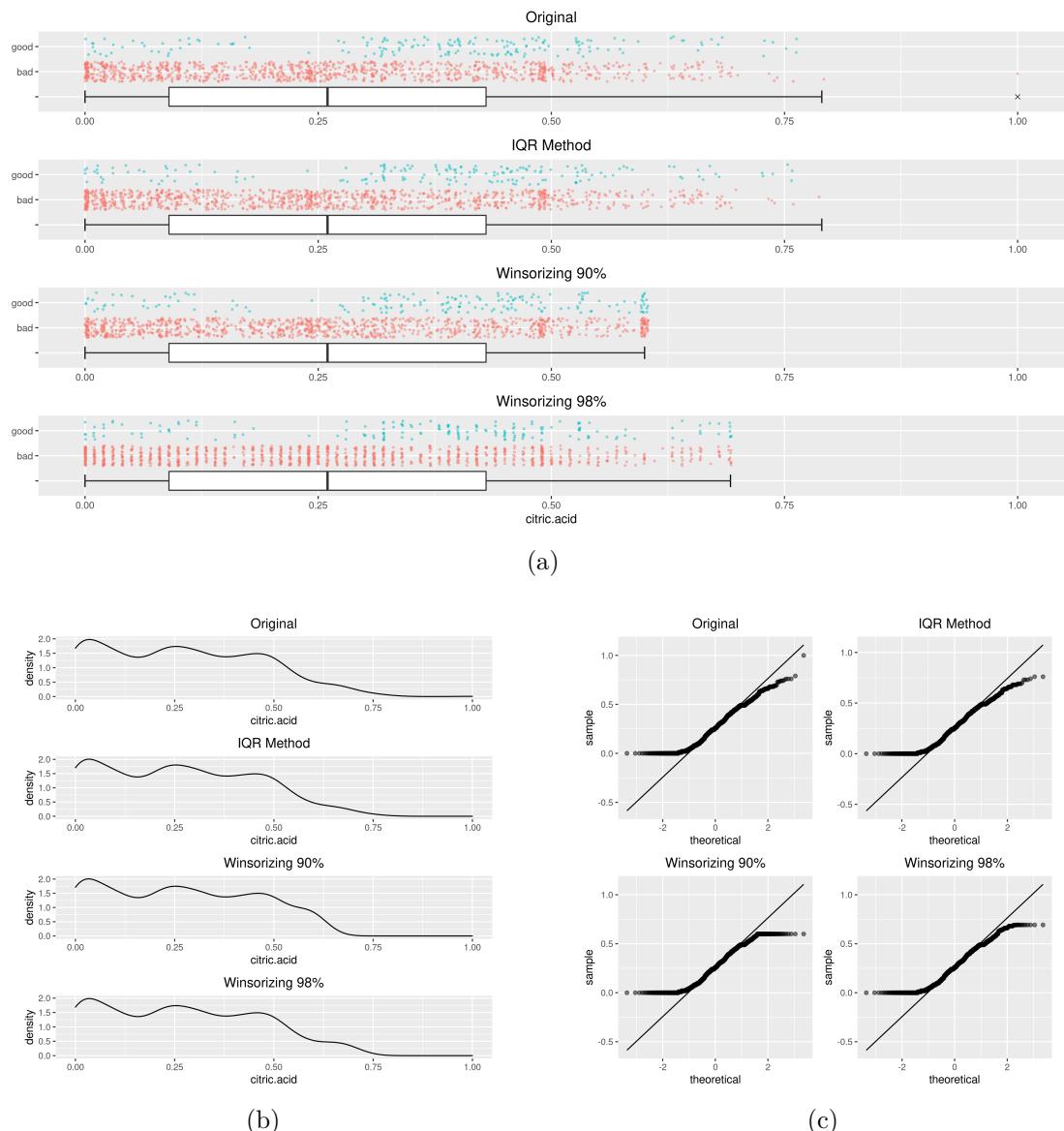
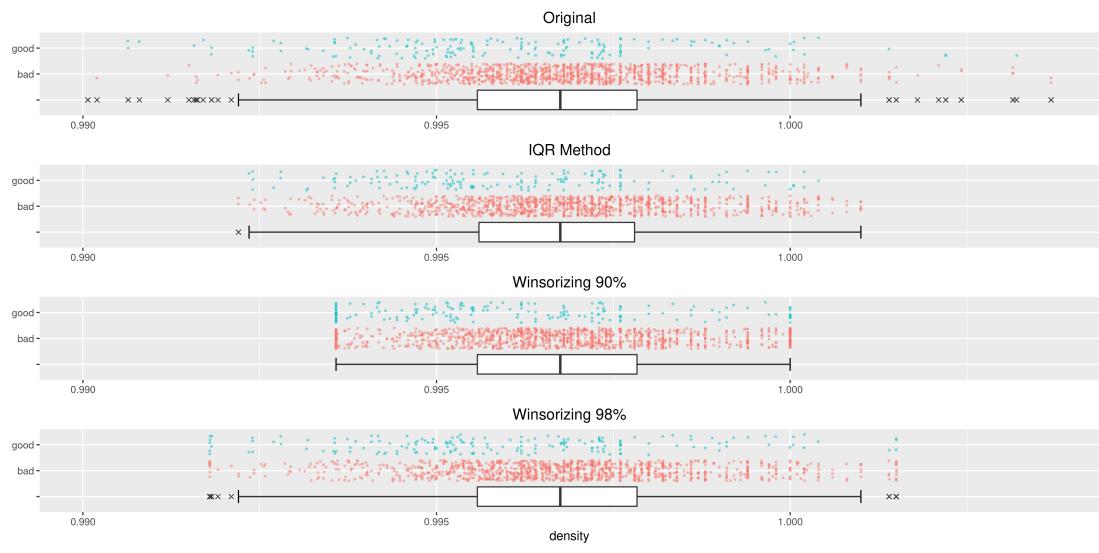
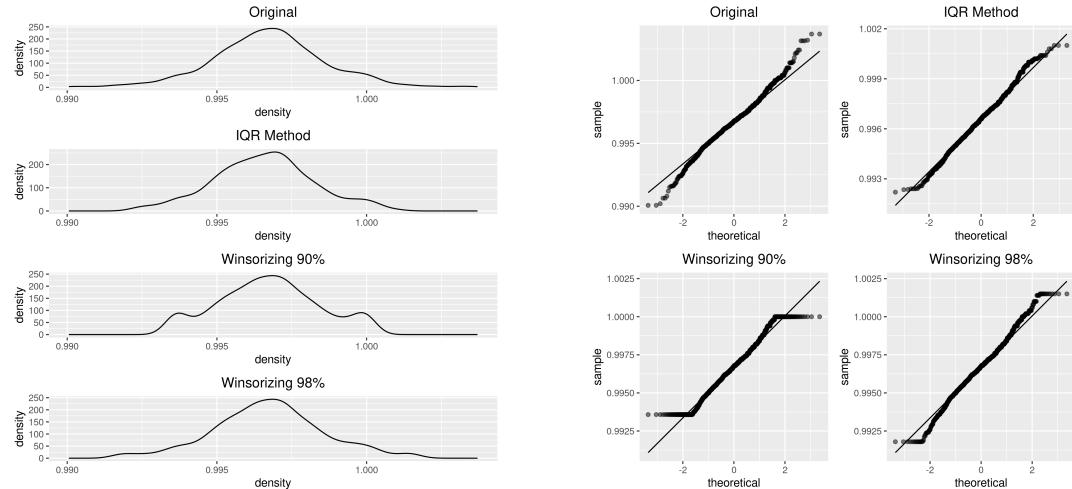


Figura 3.7: Citric Acid

Analisi Esplorativa



(a)

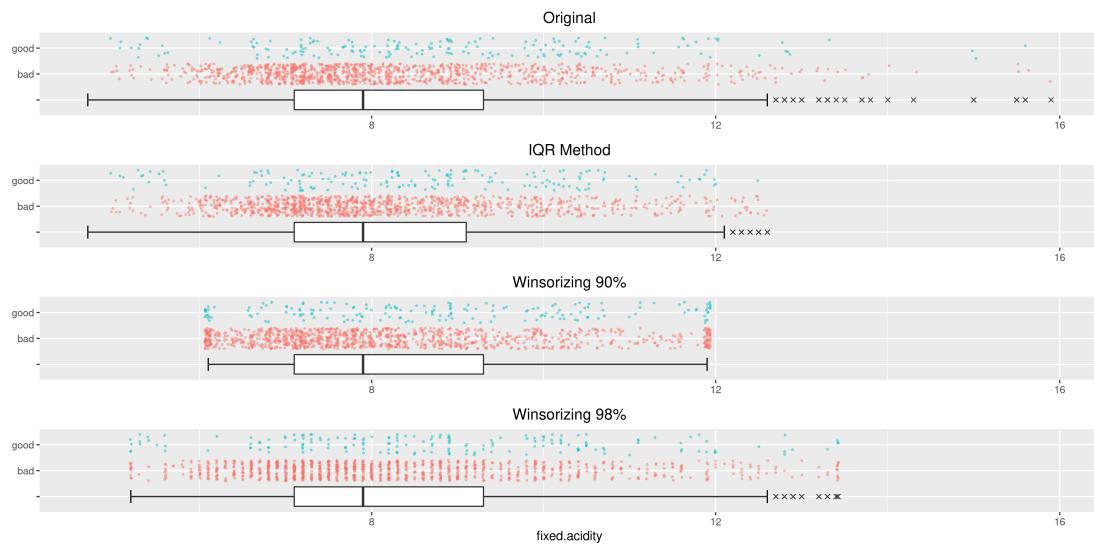


(b)

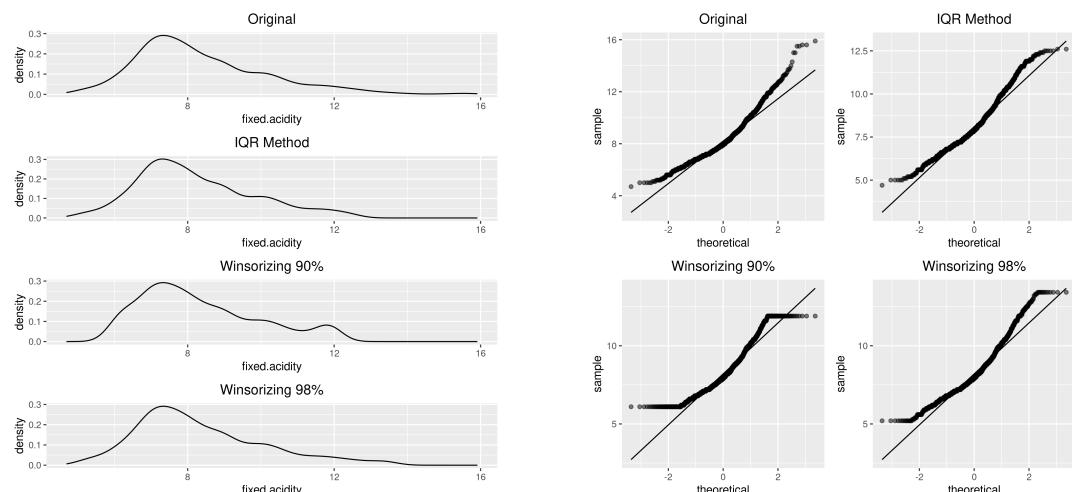
(c)

Figura 3.8: Density

Analisi Esplorativa



(a)

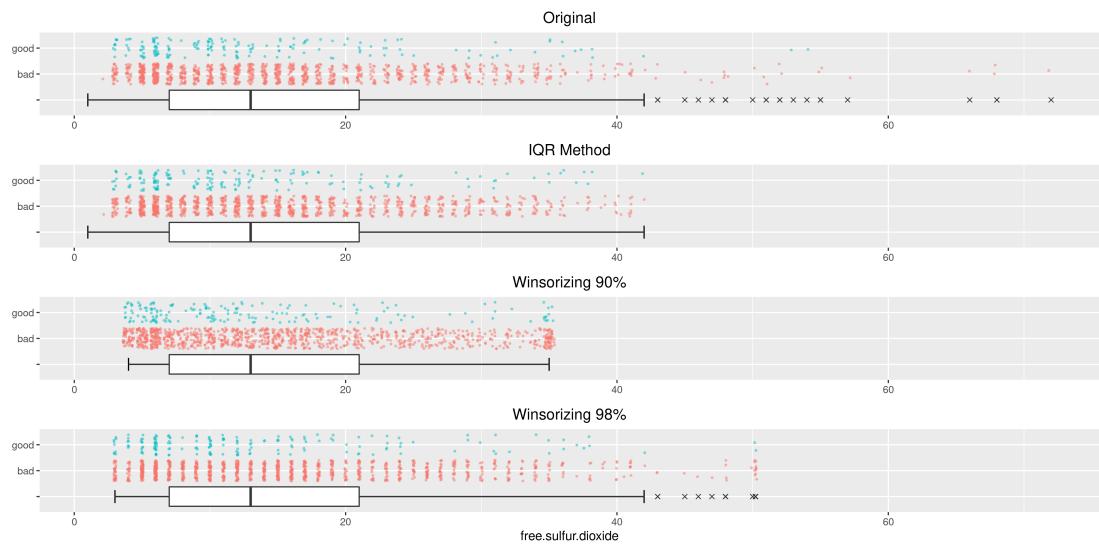


(b)

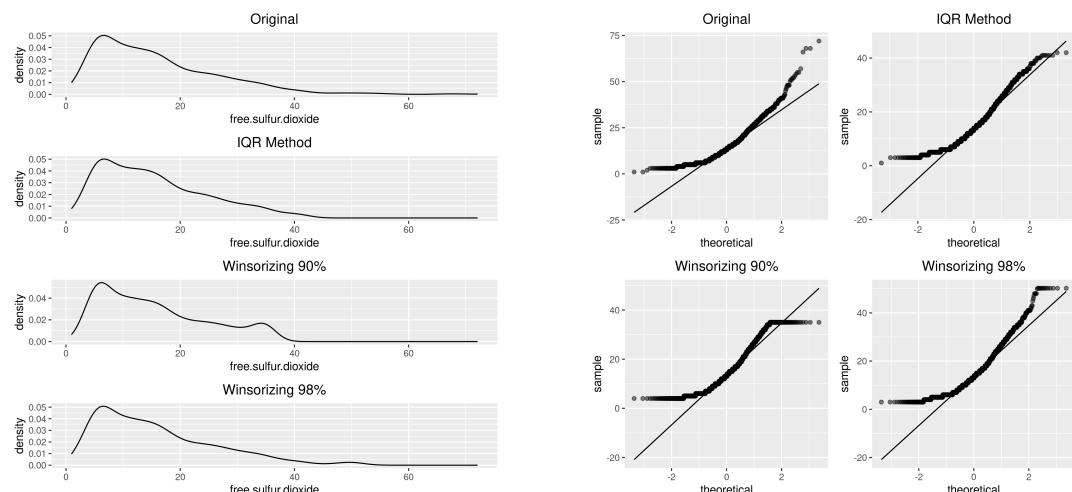
(c)

Figura 3.9: Fixed Acidity

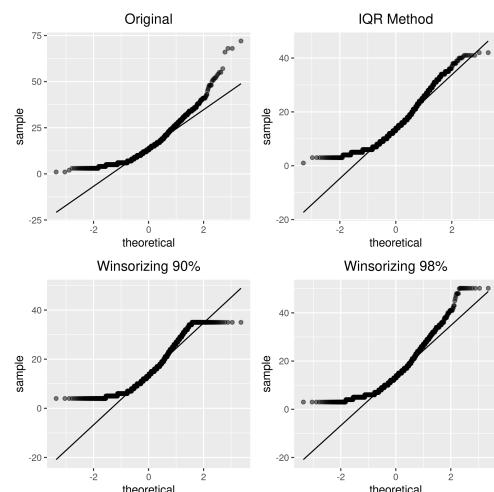
Analisi Esplorativa



(a)



(b)



(c)

Figura 3.10: Free Sulfur Dioxide

Analisi Esplorativa

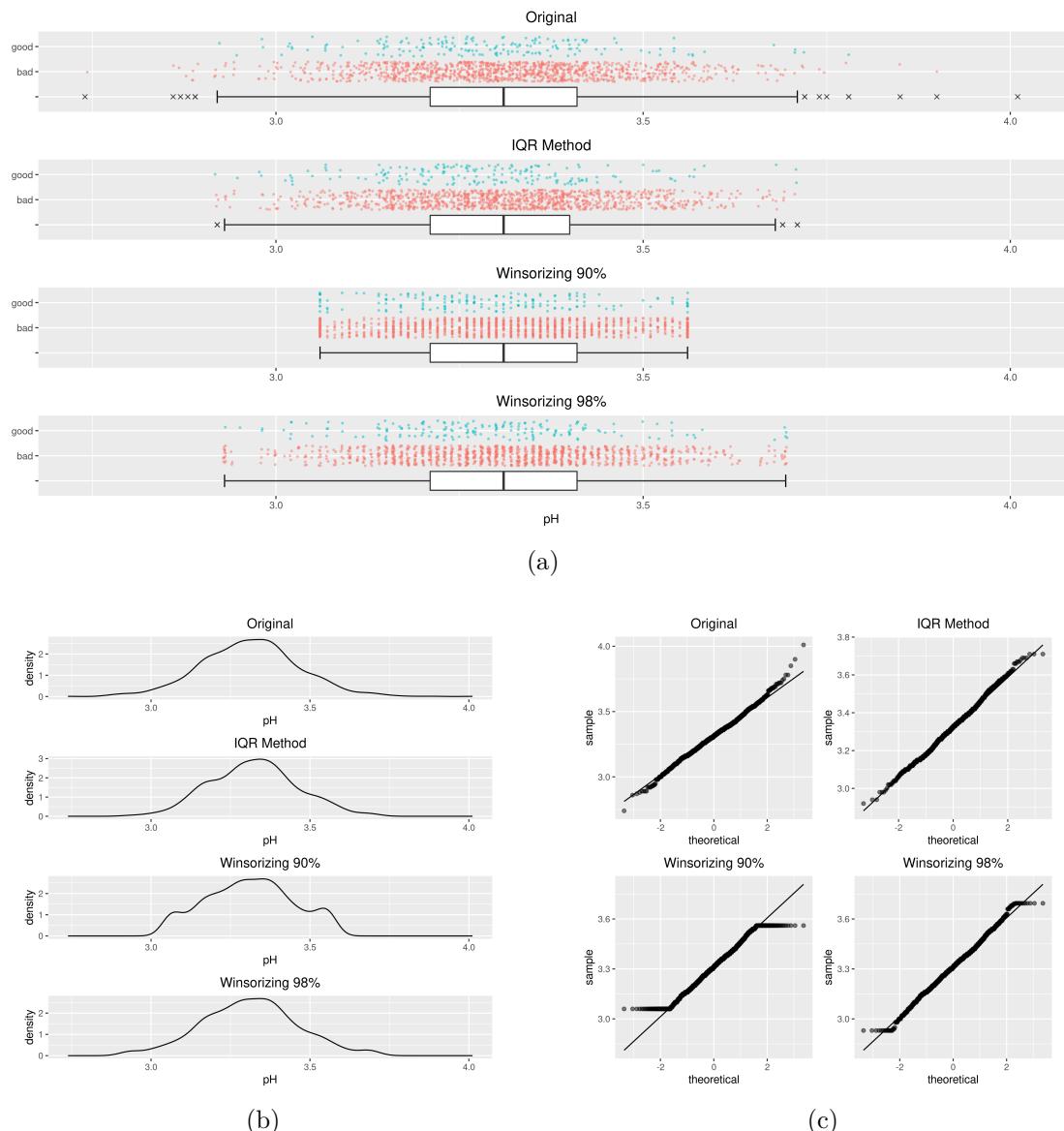


Figura 3.11: PH

Analisi Esplorativa

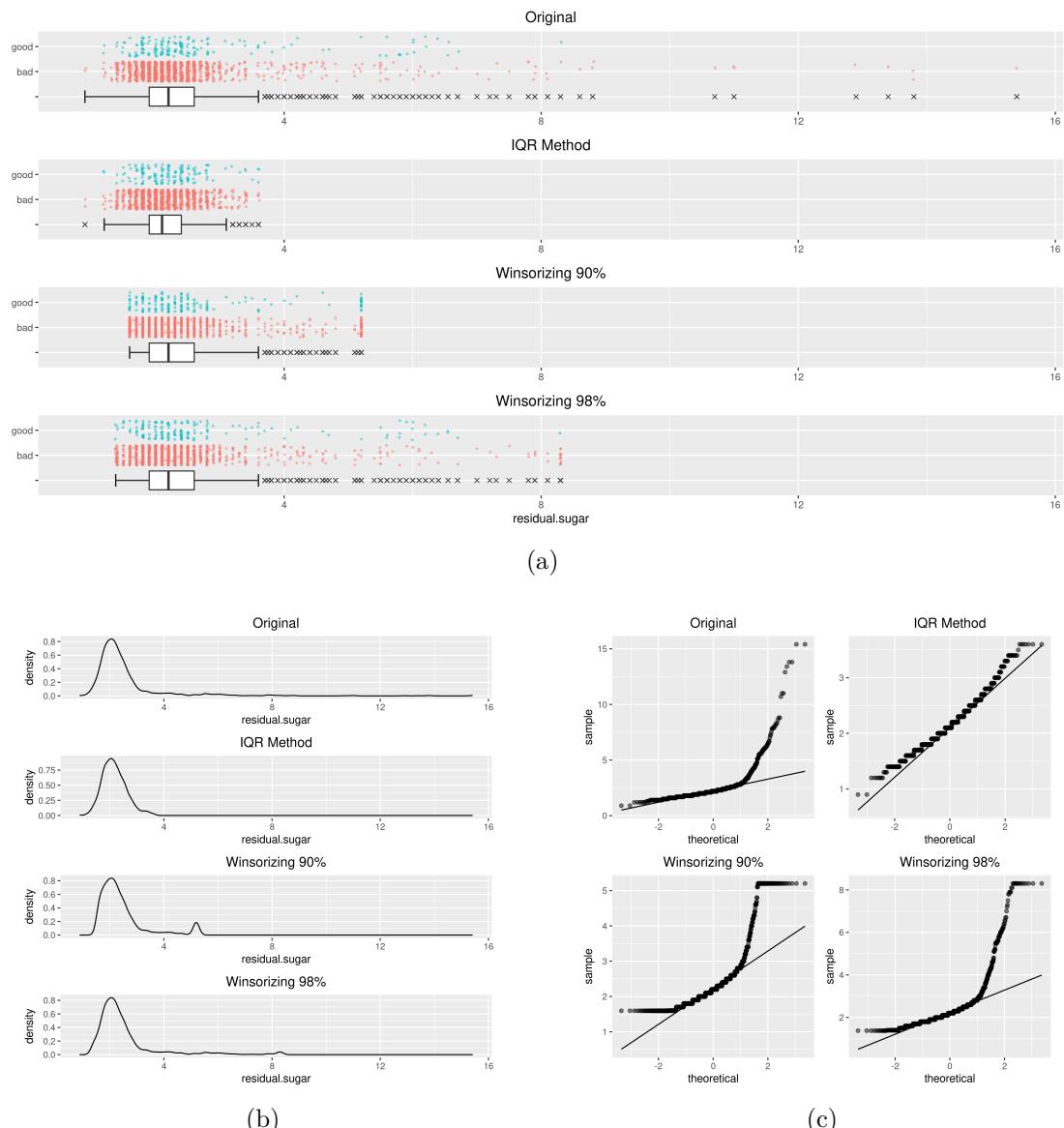
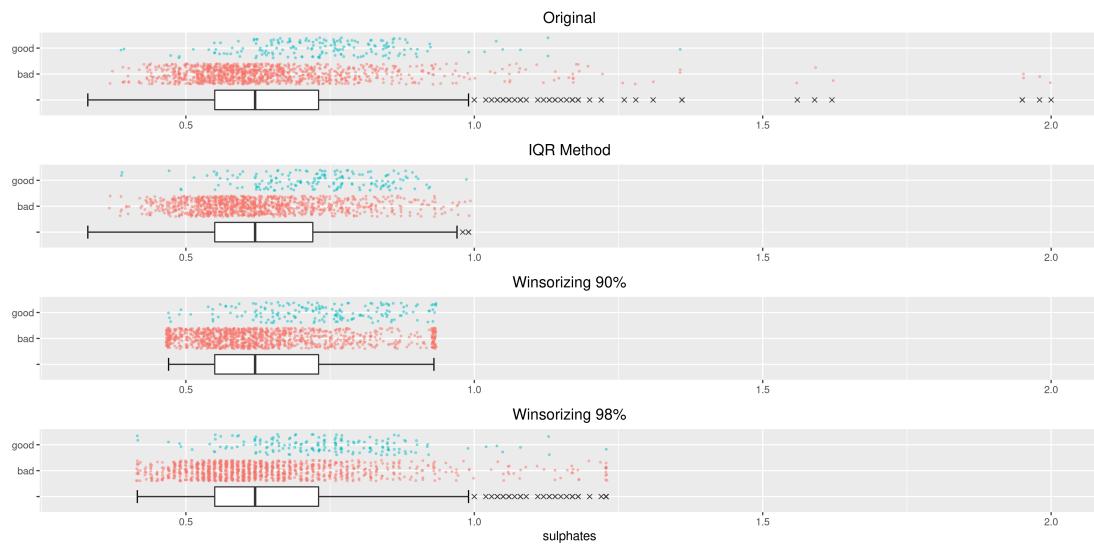
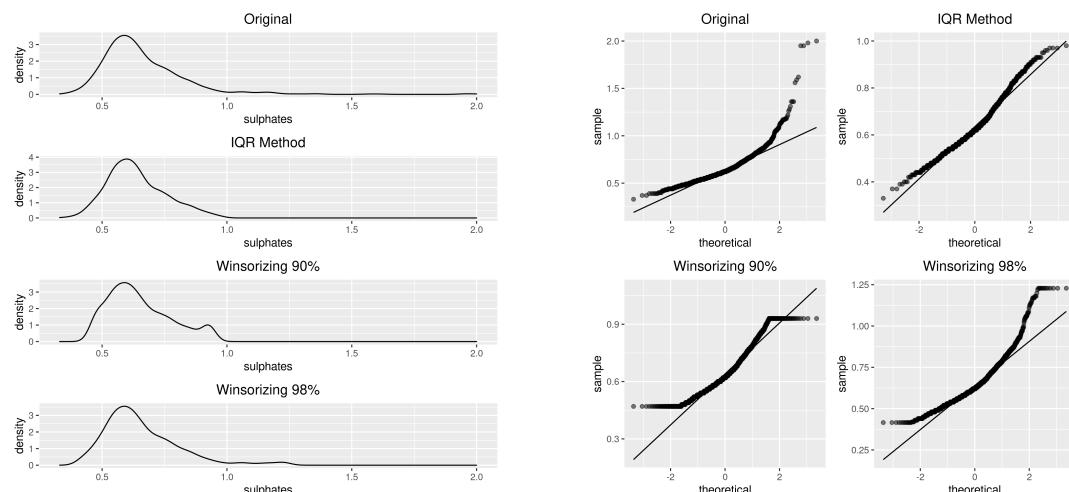


Figura 3.12: Residual Sugar



(a)

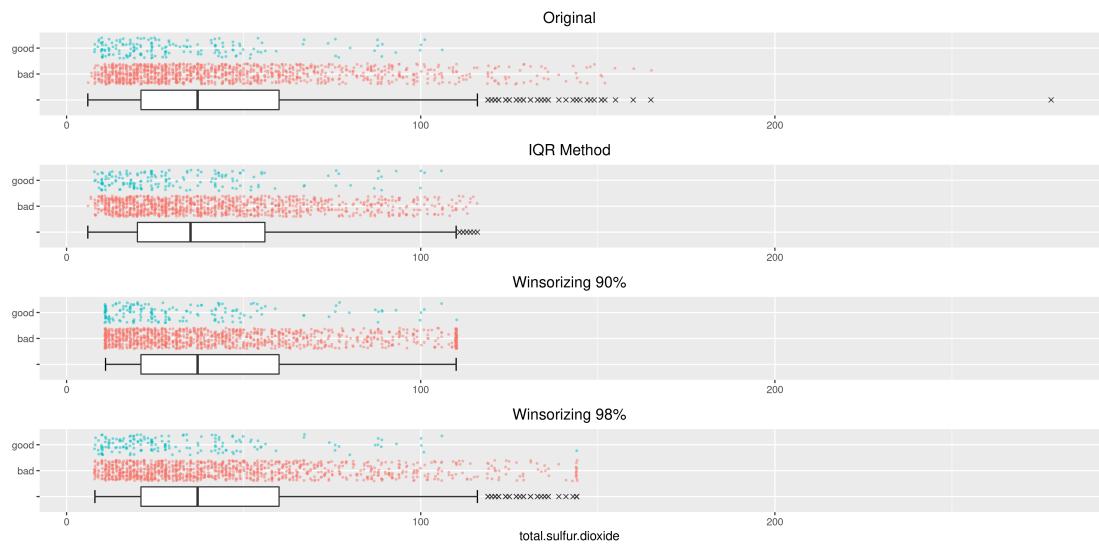


(b)

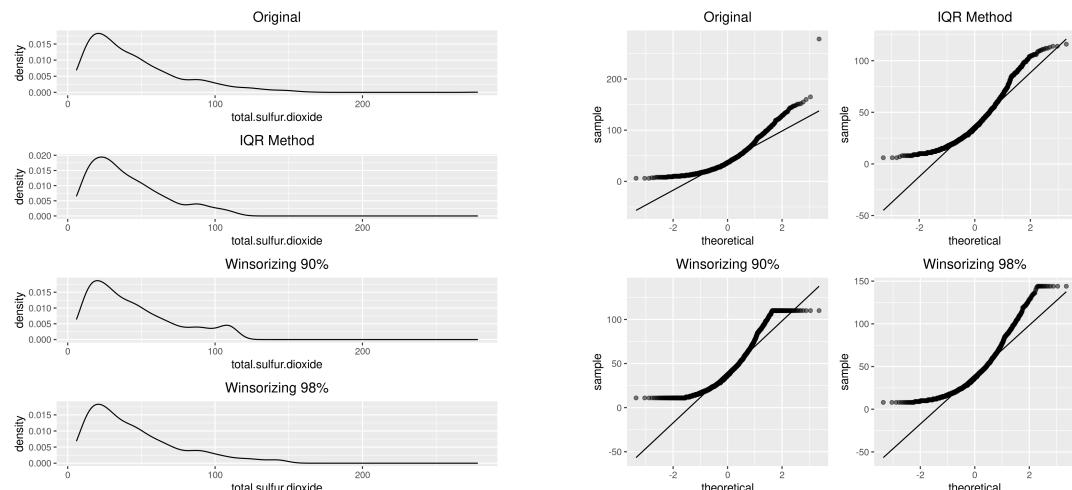
(c)

Figura 3.13: Sulphates

Analisi Esplorativa



(a)

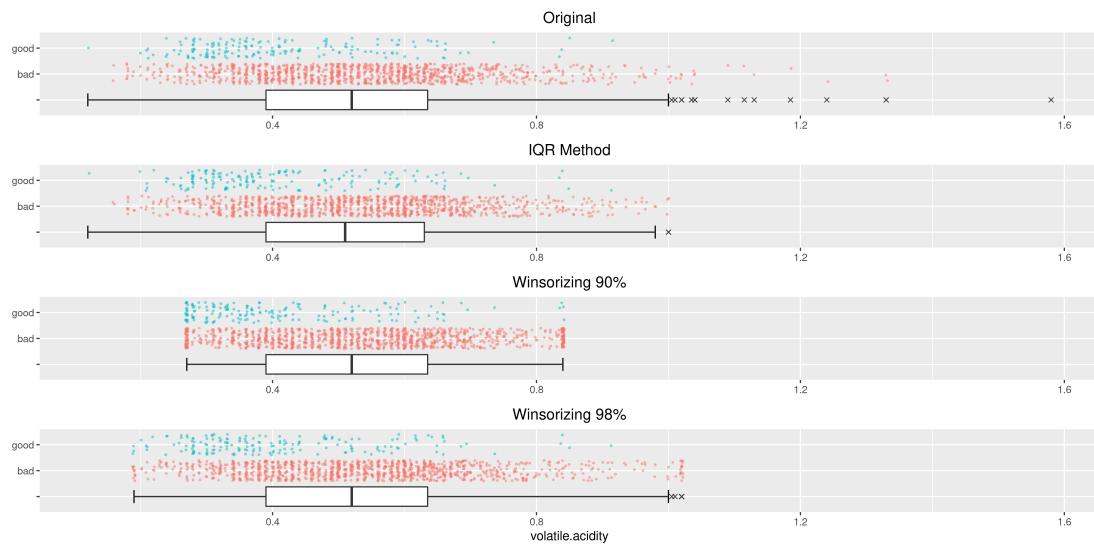


(b)

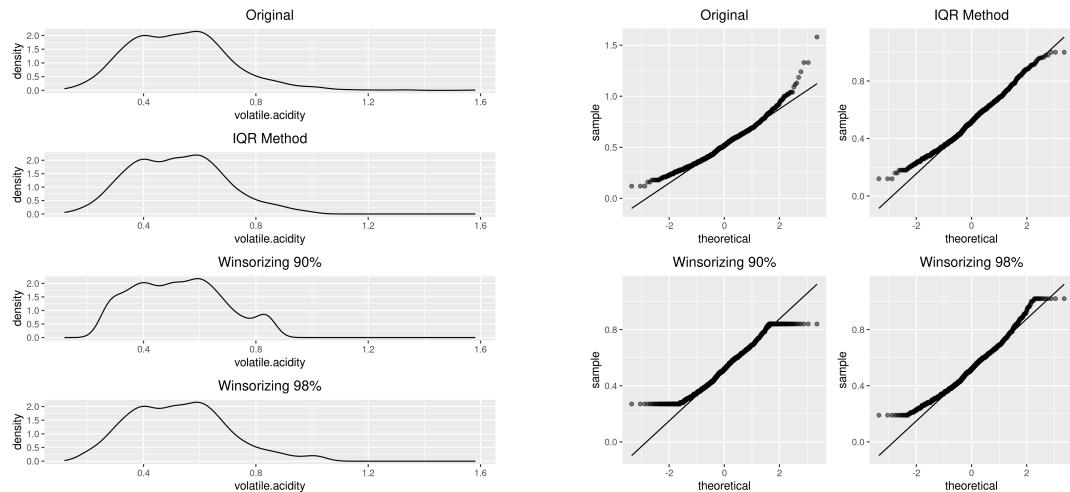
(c)

Figura 3.14: Total Sulfur Dioxide

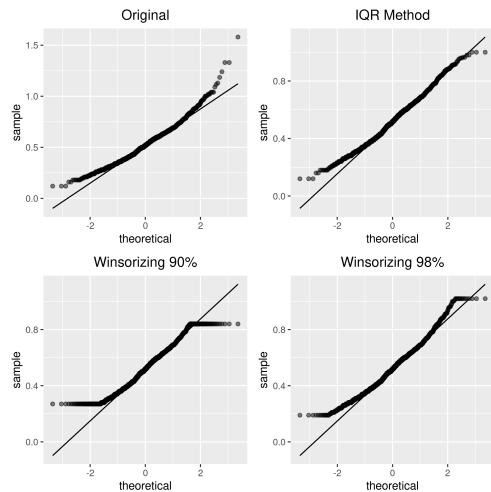
Analisi Esplorativa



(a)



(b)



(c)

Figura 3.15: Volatile Acidity

3.4 Correlazione

In questo capitolo si analizza la relazione presente tra le diverse variabili tramite la matrice di correlazione.

La matrice di correlazione mette in mostra la correlazione tra ogni coppia di variabili del dataset, per questo motivo otteniamo una matrice simmetrica.

In questo caso abbiamo sull'anti-diagonale l'incrocio con la stessa identica variabile e questo comporta la correlazione massima.

In seguito sono riportate le matrici di correlazione [3.16], [3.17] dove la seconda matrice mostra i dati dopo aver tolto gli outliers dal dataset .

I valori positivi di correlazione indicano che all'aumentare dei valori di una variabile aumentano anche i valori assunti dall'altra variabile, mentre i valori di correlazione negativi indicano che al crescere dei valori di una variabile corrisponde un andamento di decrescita nei valori dell'altra variabile.

Come si può notare otteniamo una correlazione molto bassa, questo conferma quanto precedentemente affermato riguardo alla difficoltà per un modello di poter inferire sui dati.

Osservando i valori di correlazione che caratterizzano la variabile *quality* si può notare come solo *alcohol* e *volatile.acidity* forniscono una correlazione utile per poter distinguere la classe di qualità, ma comunque con valori medio bassi e quindi poco significativi.

Inoltre questa bassa correlazione suggerisce che l'utilizzo di una PCA non porterà a miglioramenti significativi.

Come si può notare osservando i due grafici la rimozione degli outliers porta a piccoli miglioramenti aumentando la correlazione in valore assoluto.

Analisi Esplorativa

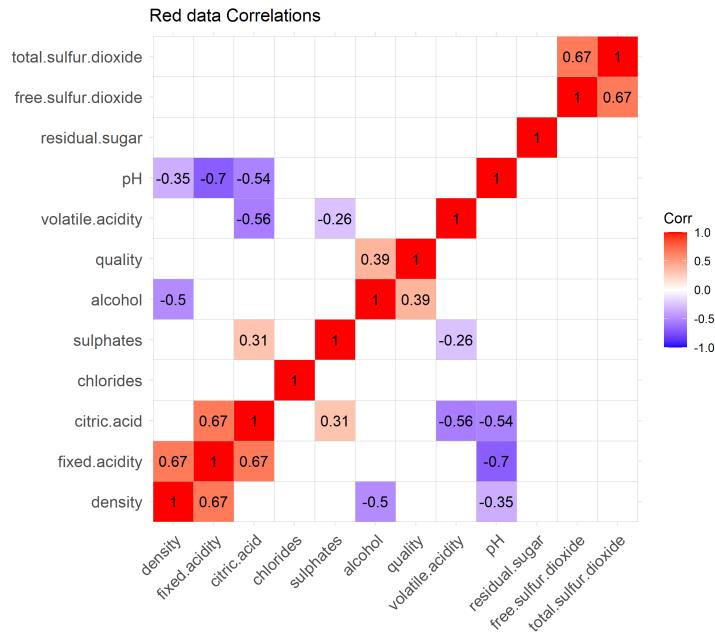


Figura 3.16: Questa immagine rappresenta una matrice della correlazione che mette in evidenza le maggiori correlazioni tra le diverse variabili.

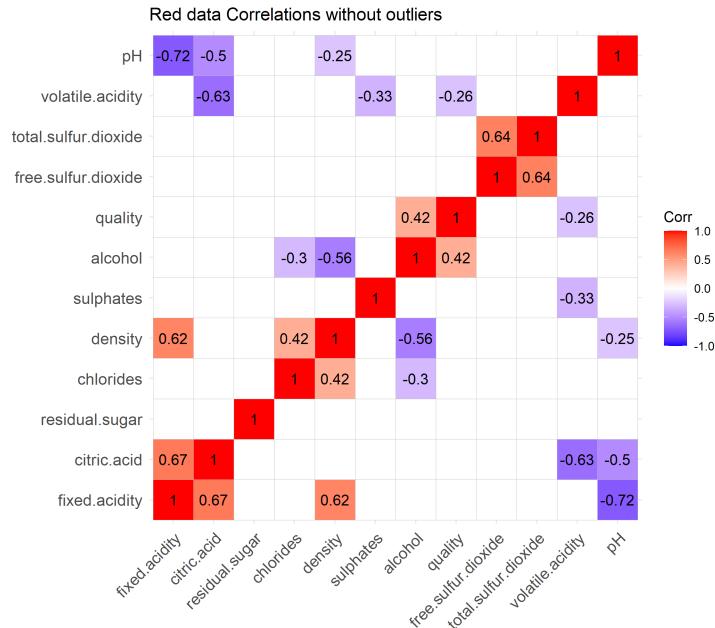


Figura 3.17: Questa immagine rappresenta una matrice della correlazione che mette in evidenza le maggiori correlazioni tra le diverse variabili, in questo specifico caso il dataset non contiene gli outlier.

3.5 Analisi delle componenti principali

L'analisi delle componenti principali [9] (in inglese principal component analysis o abbreviata PCA) è una tecnica di riduzione della dimensionalità di un insieme di dati utilizzata nell'ambito della statistica multivariata.

Lo scopo della PCA è quello di ridurre il numero più o meno elevato di variabili che descrivono un insieme di dati a un numero minore di variabili, mantenendo le più importanti e limitando il più possibile la perdita di informazioni.

Ciò avviene tramite una trasformazione lineare delle variabili che proietta quelle originarie in un nuovo sistema cartesiano [10].

La riduzione della complessità avviene limitandosi ad analizzare le componenti principali, per varianza, tra le nuove variabili ottenute e scartando quelle con poca varianza.

Nel nostro caso è stata applicata una *feature extraction* ovvero sono state estratte le componenti principali dal dominio trasformato, invece la *feature selection* seleziona un sottoinsieme delle variabili originali.

La differenza principale è che la *feature extraction* crea delle nuove variabili non presenti nel dominio originale.

Nei grafici [3.18] e [3.19] che seguono è riportata un'analisi della varianza spiegata ovvero della varianza che ogni componente rappresenta rispetto alla varianza totale. Questo grafico permette di capire se la trasformazione può portare a dei benefici e mostra la varianza spiegata da ogni variabile.

L'obiettivo è quello di prendere un numero ridotto di componenti cercando di rappresentare il 90-95% della varianza spiegata.

Come si può notare dai grafici per poter spiegare il 95% sarebbe necessario prendere almeno 8 componenti rispetto alle 11 totali; questo mostra come la PCA porta miglioramenti non molto significativi.

Anche rimuovendo gli outliers non si ottiene nessuna miglioria.

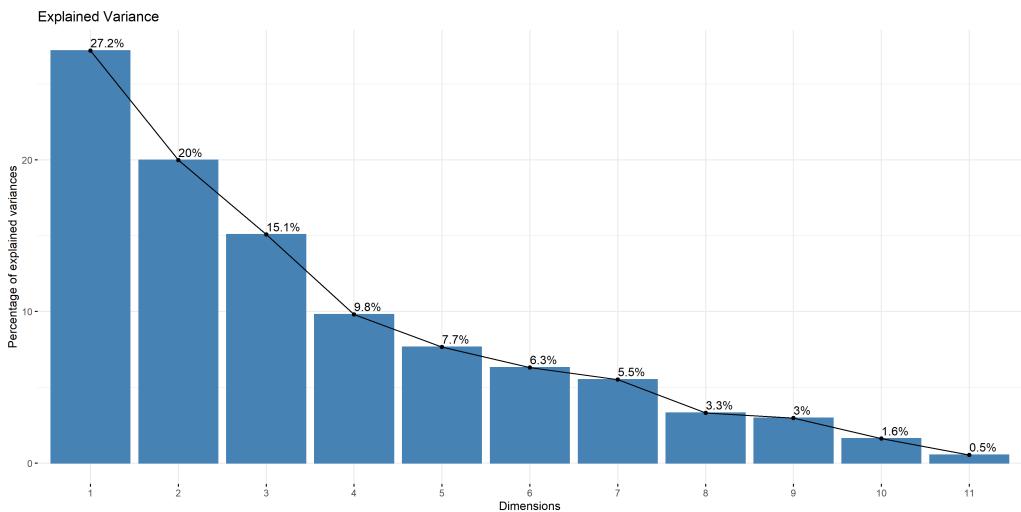


Figura 3.18: Questa immagine rappresenta la varianza spiegata per ogni componente della PCA.

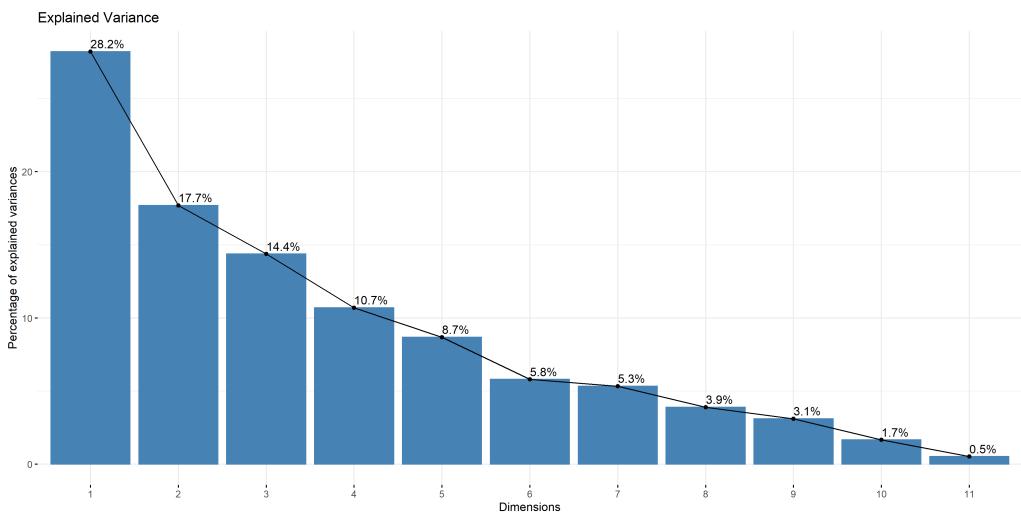


Figura 3.19: Questa immagine rappresenta la varianza spiegata per ogni componente della PCA; in questo caso dal dataset sono stati rimossi gli outliers.

Capitolo 4

Pre Processing

Il pre processing dei dati è una fase importante che consiste nel manipolare i dati attraverso varie trasformazioni e scelte. Questa fase include la rimozione dei valori mancanti, un'eventuale scelta delle istanze da usare (campionamento, rimozione degli outliers), rimozione della ridondanza, trasformazioni sui dati come ad esempio normalizzazione, standardizzazione, *feature extraction* e *feature selection*. Queste operazioni permettono di avere un input che passato ai modelli produce risultati migliori.

Il dataset originale è stato diviso in due sottoinsiemi, tenendo in ognuno di questi la stessa percentuale di istanze per classe (bad, good). Il training set contiene l'80% del dataset, mentre il test set contiene il restante 20%.

Sul training set sono state applicate le seguenti strategie di pre processing:

- Standardizzazione
- Standardizzazione + PCA
- Standardizzazione e rimozione outliers
- Standardizzazione + PCA e rimozione outliers

Pre Processing

Per la standardizzazione vengono calcolate media μ e deviazione standard σ per ogni variabile del training set, mentre per la PCA viene calcolata la matrice di rotazione W . Queste misure vengono usate per effettuare il pre processing su entrambi i dataset X (training e test). La rimozione degli outliers viene effettuata attraverso il metodo scelto (IQR), prima di applicare le varie trasformazioni sui dati.

$$Z = \frac{X - \mu}{\sigma} \quad (\text{Standardizzazione})$$

$$Z = XW \quad (\text{Trasformazione PCA})$$

Capitolo 5

Modelli

In questo breve capitolo, oltre a presentare i modelli usati: CART e SVM [5], con le loro caratteristiche principali, verranno inoltre mostrati i rispettivi parametri di tuning scelti tramite grid search per i diversi tipi di pre processing sul dataset.

5.1 CART (Classification And Regression Tree)

CART è un algoritmo di classificazione supervisionato che costruisce un albero iterativamente. Ad ogni passo viene scelto l'attributo migliore secondo un criterio prestabilito e viene associato a un nodo. L'arco verso un nodo figlio rappresenta un possibile valore per quell'attributo, mentre i nodi foglia rappresentano il valore predetto per il target. E' stato scelto come criterio *gini index*.

- Sono poco soggetti agli outliers e ai valori mancanti
- Permettono di trovare una funzione non lineare
- E' un algoritmo non parametrico, quindi non richiede assunzioni, tecniche di regolarizzazione apparte
- Molto soggetto a problematiche di overfitting se l'albero è profondo
- Sono poco costosi computazionalmente rispetto a SVM e Reti Neurali
- Non è necessario nessun pre processing al contrario delle SVM e delle Reti Neurali

Tuning

Per CART la grid search prevede come un unico parametro ovvero la profondità massima dell'albero. Tramite la scelta di questo parametro è possibile ottenere un albero meno profondo e ridurre il rischio di overfitting.

- **CART: max depth:** da 2 a 30

pre processing	max depth
Standardizzazione	14
Standardizzazione + PCA	13
Standardizzazione senza outliers	9
Standardizzazione + PCA senza outliers	9

Tabella 5.1: Parametri migliori per CART per ogni pre processing

5.2 SVM (Support Vector Machine)

L'SVM è un algoritmo di classificazione supervisionato che ha come scopo trovare il miglior iperpiano che separa due classi, massimizzando il margine tra esse.

- Nella versione soft margin permette una certa tolleranza agli outliers
- L'utilizzo dei metodi kernel consente di gestire la non linearità dei dati
- Richiede un pre processing dei dati al contrario di CART e Naive Bayes
- Al contrario di altri modelli come Naive Bayes e CART ha un grosso costo computazionale
- Rispetto alle Reti Neurali richiede meno dati per essere trainato

Tuning

Per l'SVM sono stati usati diversi kernel: lineare, polinomiale e radiale. Per ognuno di essi vengono riportati i parametri su i quali viene effettuata la grid search.

models	C	sigma	degree	scale
lineare	range	\	\	\
polinomiale	range	\	2,3,4,5,6	1/numero di features
radiale	range	range	\	\

Tabella 5.2: Parametri utilizzati per la grid search per i vari metodi kernel, range = (0.01, da 0.1 a 1.5 con passo 0.1, 2, 5, 10)

pre processing	C
Standardizzazione	1.1
Standardizzazione + PCA	0.4
Standardizzazione senza outliers	1
Standardizzazione + PCA senza outliers	1.2

Tabella 5.3: Parametri migliori per la SVM con kernel lineare per ogni pre processing

pre processing	C	sigma
Standardizzazione	0.01	2
Standardizzazione + PCA	0.01	2
Standardizzazione senza outliers	2	1.4
Standardizzazione + PCA senza outliers	0.4	1.4

Tabella 5.4: Parametri migliori per la SVM con kernel radiale per ogni pre processing

pre processing	C	scale	degree
Standardizzazione	0.4	0.09090	2
Standardizzazione + PCA	0.01	0.11111	2
Standardizzazione senza outliers	0.01	0.09090	2
Standardizzazione + PCA senza outliers	0.1	0.11111	2

Tabella 5.5: Parametri migliori per la SVM con kernel polinomiale per ogni pre processing

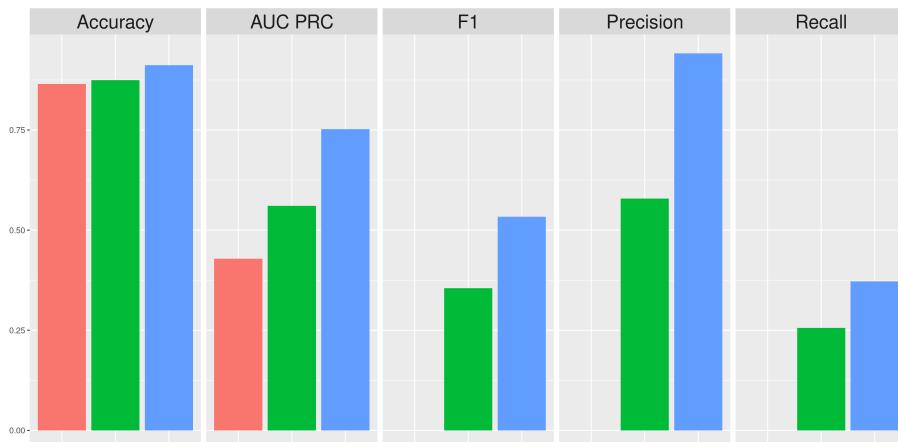
Capitolo 6

Esperimenti

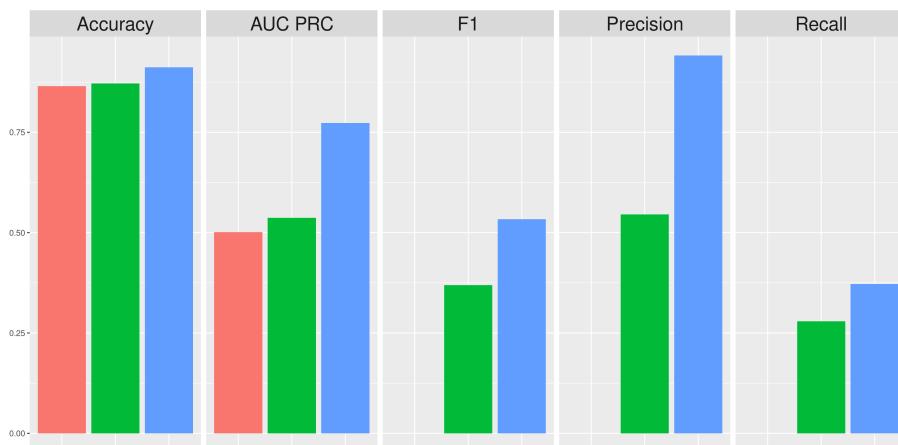
In questo capitolo vengono mostrati i vari esperimenti fatti in base al pre processing utilizzato. I vari modelli vengono addestrati utilizzando una k-fold cross validation. La cross validation necessita di essere eseguita in modo appropriato poiché il dataset in esame è sbilanciato. Vista la numerosità della classe minoritaria si è fissato $k=5$ rispetto al solito valore di $k=10$ in modo da avere dei sottoinsiemi significativi. Inoltre la cross validation viene effettuata in modo stratificato, così da avere in ogni fold in proporzione lo stesso numero di istanze positive e negative, poiché altrimenti potrebbe accadere che un fold abbia pochi elementi della classe minoritaria. La cross validation viene effettuata 5 volte e alla fine viene scelto il modello che massimizza l'AUC (Area Under The Curve) della PRC (Precision Recall Curve). Questa metrica è comunemente utilizzata in caso di dati sbilanciati, l'accuracy in questi casi potrebbe portare a risultati errati, poiché si concentra prevalentemente sulla classe maggioritaria. I vari esperimenti verranno mostrati di seguito nel seguente ordine, prima verranno confrontate i vari metodi kernel per l'SVM, dopodiché l'SVM migliore verrà confrontata con CART, infine, dopo aver scelto il pre processing migliore, verranno confrontati i due modelli migliori.

6.1 Confronto tra i kernel per SVM

I seguenti istogrammi comparano le performance ottenute tramite i diversi kernel utilizzati per i diversi tipi di pre processing utilizzati. Le metriche adottate sono Accuracy, AUC PRC, F1, Precision, Recall. Da essi è possibile notare come con il kernel radiale si assumono valori più alti del kernel polinomiale in presenza di outliers. Rimuovendo gli outliers invece, il kernel polinomiale supera il kernel radiale in F1 e Recall. Dai risultati dell'SVM lineare si può pensare che i dati non siano linearmente separabili.

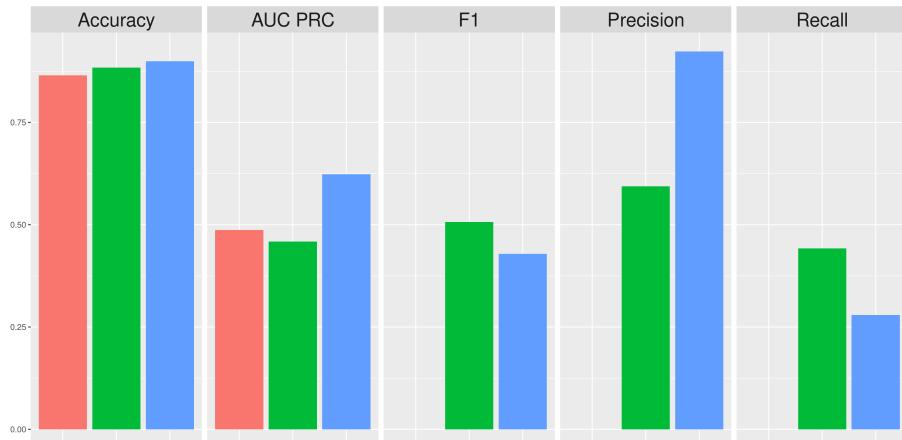


(a) Standardizzazione

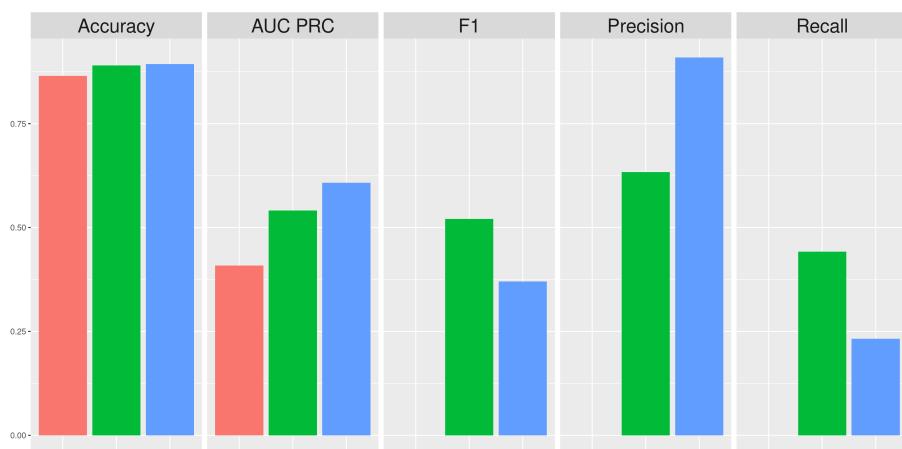


(b) Standardizzazione + PCA

Figura 6.1: Risultati della SVM con i diversi kernel (lineare: rosso, polinomiale: verde, radiale: blu) sul testset con outliers



(a) Standardizzazione



(b) Standardizzazione + PCA

Figura 6.2: Risultati della SVM con i diversi kernel (lineare: rosso, polinomiale: verde, radiale: blu) sul testset senza outliers

Mostriamo ora graficamente la curva ROC (Receiver Operating Characteristic) e la PRC (Precision Recall Curve) che è più sensibile a situazioni con dati sbilanciati. Infatti si può notare come le differenze tra i vari modelli siano molto grandi rispetto alla PRC, invece rispetto alla ROC tutti i modelli sono molto simili. Inoltre la rimozione degli outliers comporta una perdita drastica nelle performance dei modelli.

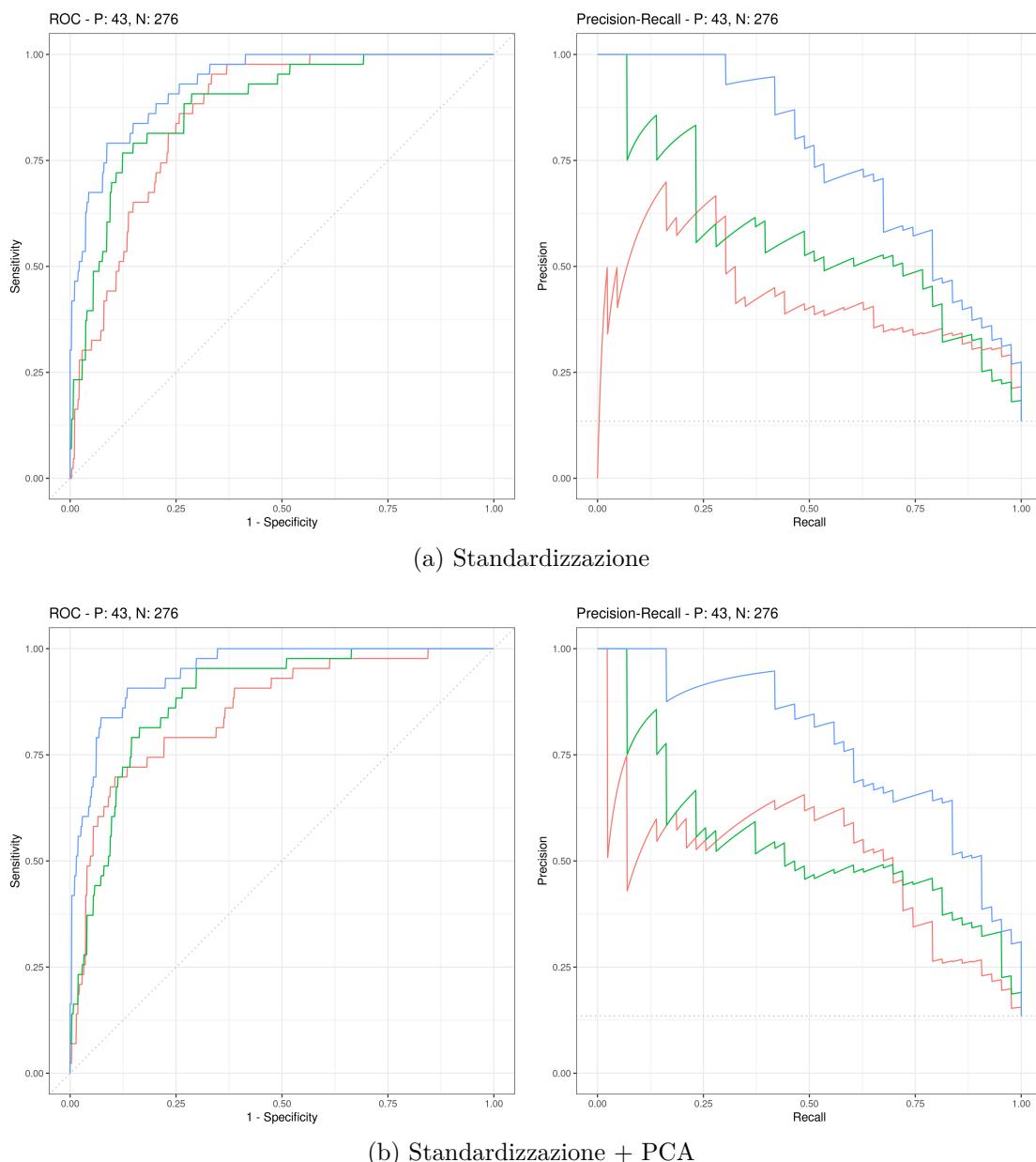


Figura 6.3: Curve ROC e PRC per i diversi kernel (lineare: rosso, polinomiale: verde, radiale: blu) sul testset con outliers

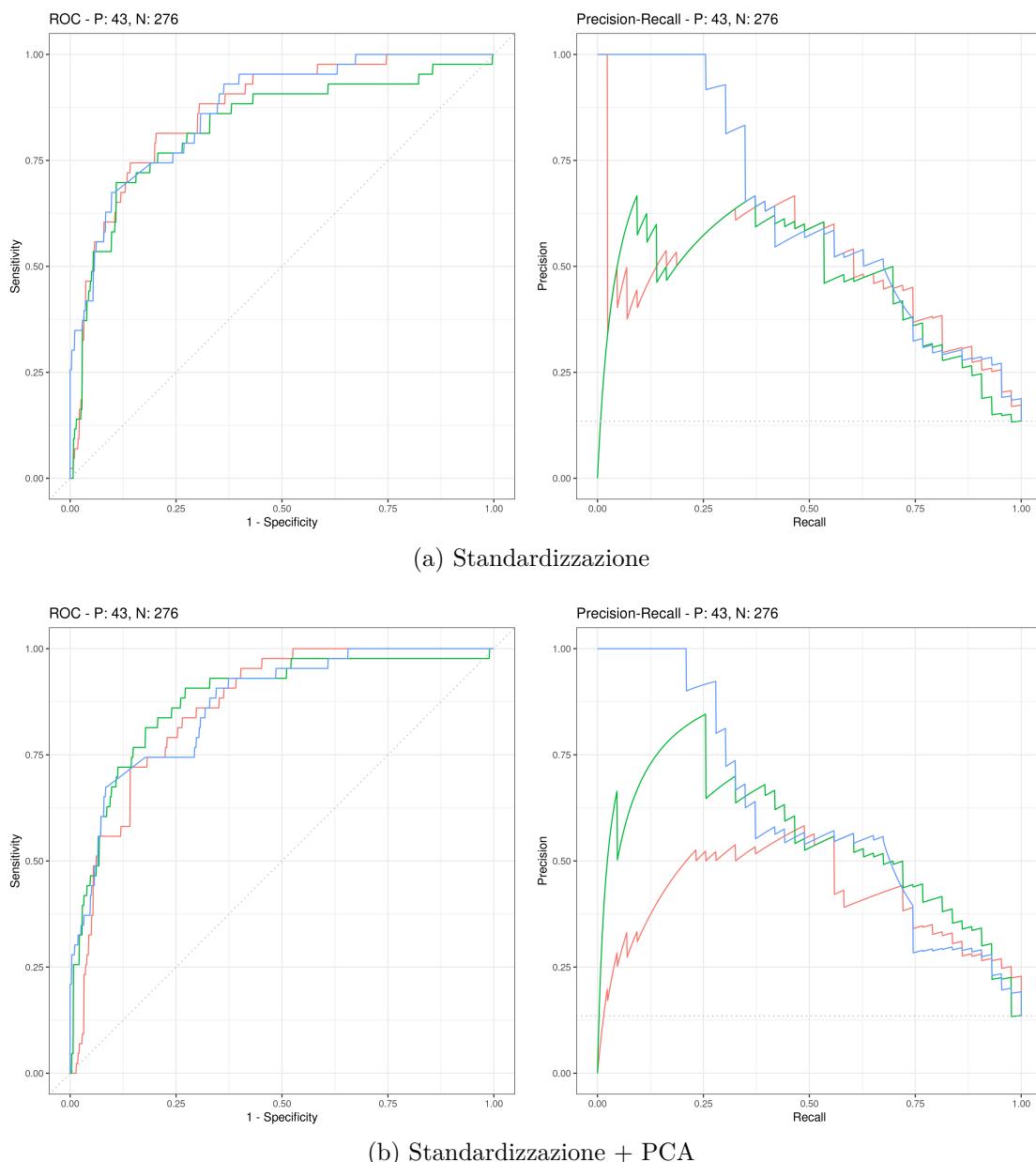


Figura 6.4: Curve ROC e PRC per i diversi kernel (lineare: rosso, polinomiale: verde, radiale: blu) sul testset senza outliers

Riassumendo nella seguente tabella è possibile vedere come il kernel radiale superi gli altri kernel mantenendo gli outliers, mentre effettuando la rimozione con il kernel polinomiale si assumono valori più alti di F1 e Recall. E' possibile notare che la ROC e la PRC si abbassano drasticamente rimuovendo gli outliers. Possiamo quindi concludere che l'SVM con kernel radiale, a prescindere dal pre processing, risulti il modello migliore nella maggioranza dei casi analizzati.

Kernel	Overall Accuracy	Precision	Recall	F1	ROC AUC	PRC AUC	95% CI	P-Value
lineare	0.8652	NA	0	NA	0.8604651	0.4288328	(0.8228, 0.9007)	0.5405
polinomiale	0.8746	0.57895	0.25581	0.35484	0.8795079	0.5608824	(0.8332, 0.9089)	0.3471558
radiale	0.9122	0.94118	0.37209	0.53333	0.9313279	0.7520759	(0.8756, 0.9409)	0.006404

Tabella 6.1: Risultati dei diversi kernel sul testset con Standardizzazione

Kernel	Overall Accuracy	Precision	Recall	F1	ROC AUC	PRC AUC	95% CI	P-Value
lineare	0.8652	NA	0	NA	0.8547354	0.5011905	(0.8228, 0.9007)	0.5405
polinomiale	0.8715	0.54545	0.27907	0.36923	0.8844793	0.5368917	(0.8297, 0.9062)	0.410227
radiale	0.9122	0.94118	0.37209	0.53333	0.9469161	0.7732596	(0.8756, 0.9409)	0.006404

Tabella 6.2: Risultati dei diversi kernel sul testset con Standardizzazione + PCA

Kernel	Overall Accuracy	Precision	Recall	F1	ROC AUC	PRC AUC	95% CI	P-Value
lineare	0.8652	NA	0	NA	0.8681328	0.4870888	(0.8228, 0.9007)	0.5405
polinomiale	0.884	0.59375	0.44186	0.50667	0.8313953	0.4587525	(0.8437, 0.917)	0.1845
radiale	0.8997	0.92308	0.27907	0.42857	0.8711662	0.6230968	(0.8613, 0.9304)	0.03864

Tabella 6.3: Risultati dei diversi kernel sul testset con Standardizzazione e rimozione outliers

Kernel	Overall Accuracy	Precision	Recall	F1	ROC AUC	PRC AUC	95% CI	P-Value
lineare	0.8652	NA	0	NA	0.8624031	0.408548	(0.8228, 0.9007)	0.5405
polinomiale	0.8903	0.63333	0.44186	0.52055	0.878244	0.5411327	(0.8507, 0.9224)	0.10722
radiale	0.8934	0.90909	0.23256	0.37037	0.8684277	0.6079154	(0.8543, 0.9251)	0.07852

Tabella 6.4: Risultati dei diversi kernel sul testset con Standardizzazione + PCA e rimozione outliers

6.2 Confronto fra CART e SVM Radiale

Così come già mostrato in precedenza per i kernel dell'SVM, vengono ora riportate le performance ottenute dall'SVM radiale e CART. Le metriche considerate sono Accuracy, AUC PRC, F1, Precision e Recall. Da essi è possibile notare come, a parte il caso [6.5a], l'SVM raggiunga i risultati migliori nella maggioranza dei casi. Nel caso [6.5a], invece è difficile dire quale dei due modelli sia il migliore dal momento che i risultati sono molto simili.

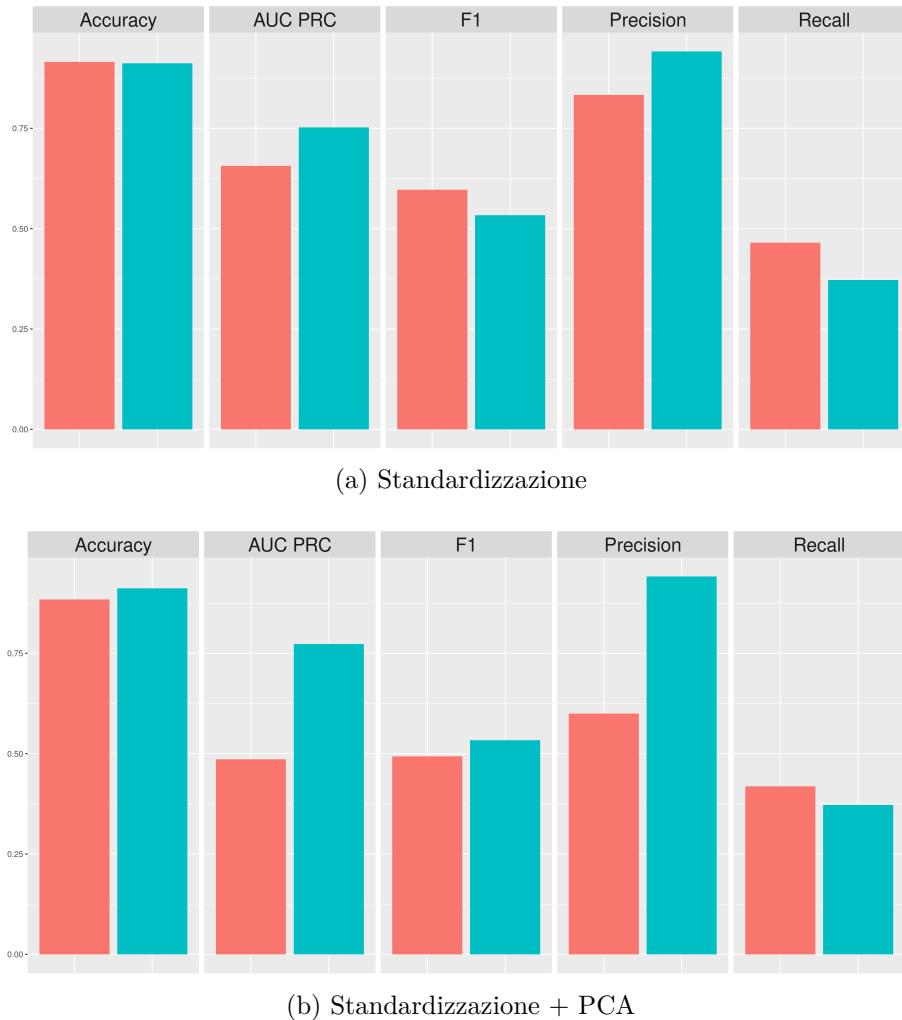
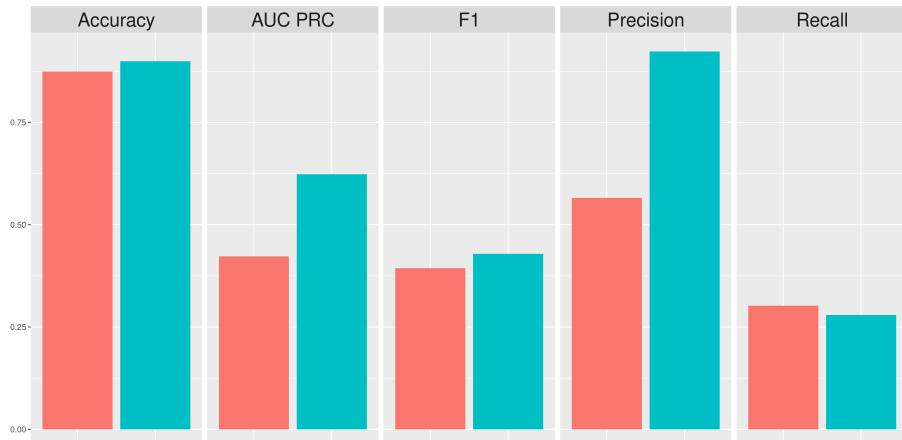


Figura 6.5: Risultati dei due modelli (cart: rosso, radiale: blu) sul testset con outliers



(a) Standardizzazione



(b) Standardizzazione + PCA

Figura 6.6: Risultati dei due modelli (cart: rosso, radiale: blu) sul testset senza outliers

Osservando le curve ROC e PRC possiamo notare come quelle dell'SVM radiale siano sempre superiori a quelle di CART. Si riconferma che la rimozione degli outliers peggiora i risultati. Dal caso [6.7a] non è possibile distinguere i risultati dalla curva PRC, per questo sono stati riportati in tabelle i valori di AUC (Area Under The Curve) che confermano che anche in questo caso l'SVM radiale raggiunge risultati migliori.

Esperimenti

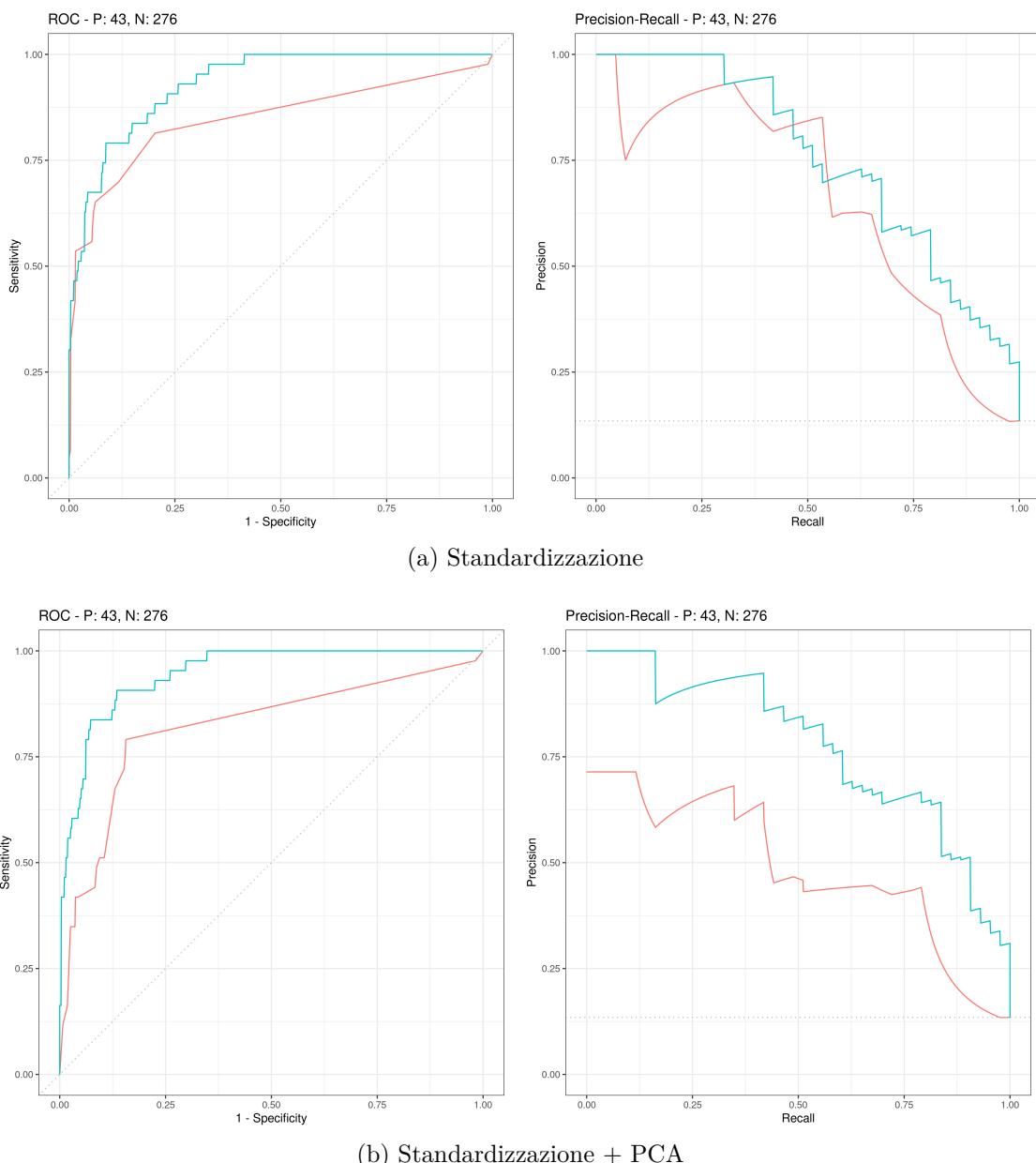


Figura 6.7: Curve ROC e PRC per i due modelli (cart: rosso, radiale: blu) sul testset con outliers

Esperimenti

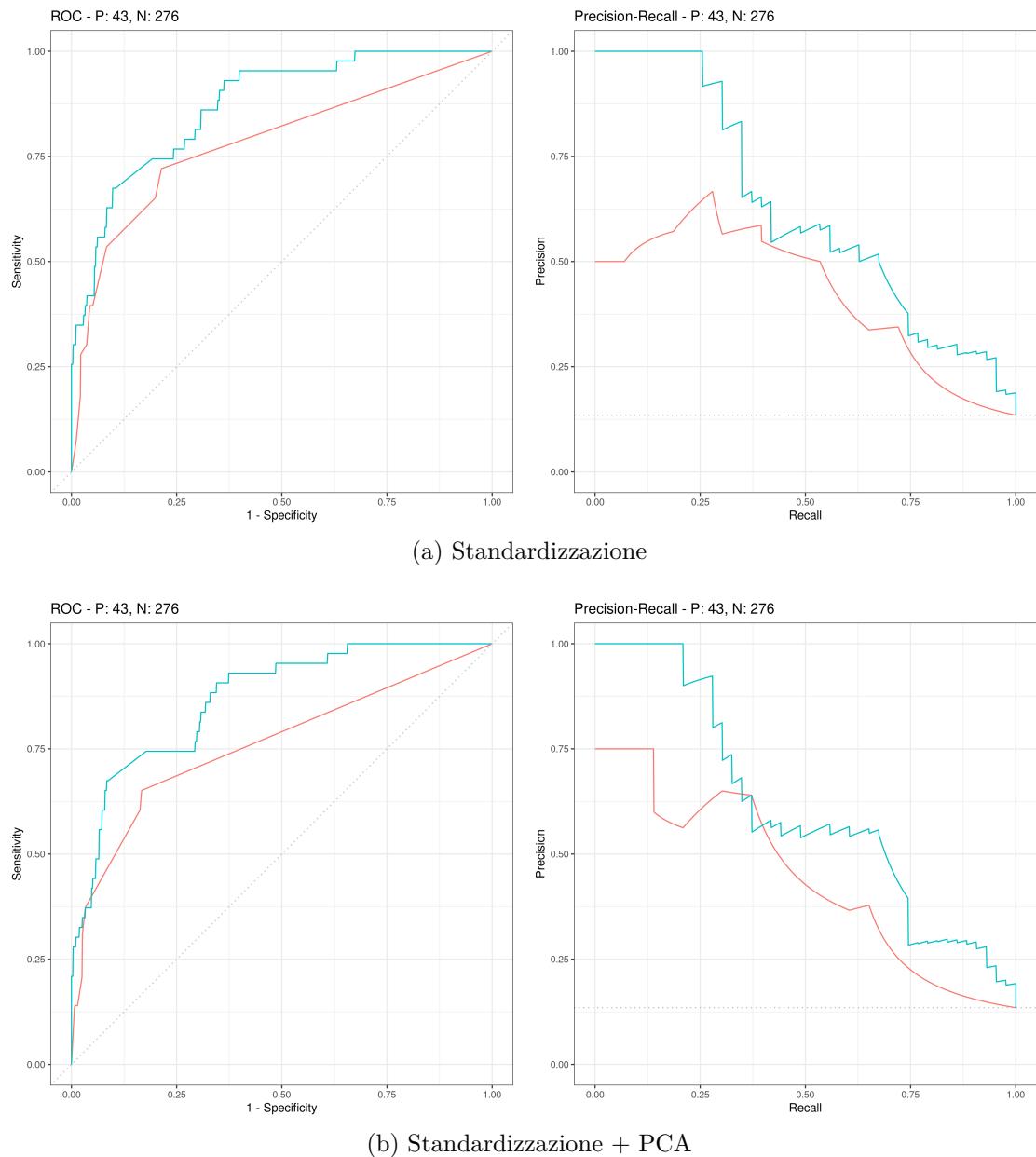


Figura 6.8: Curve ROC e PRC per i due modelli (cart: rosso, radiale: blu) sul testset senza outliers

Esperimenti

Riassumiamo nuovamente i risultati presentati nelle tabelle sottostanti. Per scegliere i modelli migliori che verranno proposti in questo lavoro, è necessario analizzare le differenze tra Standardizzazione + PCA e Standardizzazione nel caso in cui gli outliers vengono mantenuti. CART con Standardizzazione ha risultati migliori della controparte con Standardizzazione + PCA. L'SVM radiale con Standardizzazione + PCA è leggermente migliore rispetto alla controparte con Standardizzazione.

Models	Overall Accuracy	Precision	Recall	F1	ROC AUC	PRC AUC	95% CI	P-Value
cart	0.9154	0.83333	0.46512	0.59701	0.8476154	0.6564769	(0.8792, 0.9435)	0.003747
svm	0.9122	0.94118	0.37209	0.53333	0.9313279	0.7520759	(0.8756, 0.9409)	0.006404

Tabella 6.5: Risultati modelli scelti con Standardizzazione

Models	Overall Accuracy	Precision	Recall	F1	ROC AUC	PRC AUC	95% CI	P-Value
cart	0.884	0.6	0.4186	0.49315	0.8194725	0.485855	(0.8437, 0.917)	0.18452
svm	0.9122	0.94118	0.37209	0.53333	0.9469161	0.7732596	(0.8756, 0.9409)	0.006404

Tabella 6.6: Risultati modelli scelti con Standardizzazione + PCA

Models	Overall Accuracy	Precision	Recall	F1	ROC AUC	PRC AUC	95% CI	P-Value
cart	0.8746	0.56522	0.30233	0.39394	0.7817661	0.4226265	(0.8332, 0.9089)	0.347156
svm	0.8997	0.92308	0.27907	0.42857	0.8711662	0.6230968	(0.8613, 0.9304)	0.03864

Tabella 6.7: Risultati modelli scelti con Standardizzazione e rimozione outliers

Models	Overall Accuracy	Precision	Recall	F1	ROC AUC	PRC AUC	95% CI	P-Value
cart	0.8746	0.58824	0.23256	0.33333	0.7598163	0.440484	(0.8332, 0.9089)	0.3472
svm	0.8934	0.90909	0.23256	0.37037	0.8684277	0.6079154	(0.8543, 0.9251)	0.07852

Tabella 6.8: Risultati modelli scelti con Standardizzazione + PCA e rimozione outliers

Concludiamo che i modelli migliori sono CART con Standardizzazione e SVM radiale con Standardizzazione + PCA, entrambi mantenendo gli outliers.

6.3 Confronto fra i modelli scelti

In questa sezione andiamo a confrontare i due modelli proposti per la trattazione del nostro problema.

Mostriamo ora le matrici di confusione dei modelli testati sul testset composto da 319 istanze di cui 276 positive e 43 negative. Dalle matrici di confusione possiamo vedere che i modelli non sembrano così diversi e come entrambi soffrono dello sbilanciamento del dataset. Infatti i nostri modelli classificano correttamente quasi tutte le istanze appartenenti alla classe maggioritaria, mentre sbagliano molto di più sulla classe minoritaria.



Figura 6.9: Matrice di confusione sul testset per SVM Radiale con Standardizzazione + PCA (sinistra) e CART con Standardizzazione (destra)

Comparando le metriche ottenute dai due modelli possiamo vedere come non ci sia una grossa differenza, tra alti valori di Precision e bassi valori di Recall, non considerando l'Accuracy come già detto precedente non è un buon indicatore. L'AUC è un indicatore migliore infatti notiamo che l'SVM radiale supera CART in AUC e Precision, mentre CART è migliore in Recall e F1.

Esperimenti

Mostrando la ROC e la PRC è possibile vedere come l'SVM radiale superi CART in entrambe seppure di poco, rispettivamente 0.9469161/0.7732596 0.9313279/0.7520759.

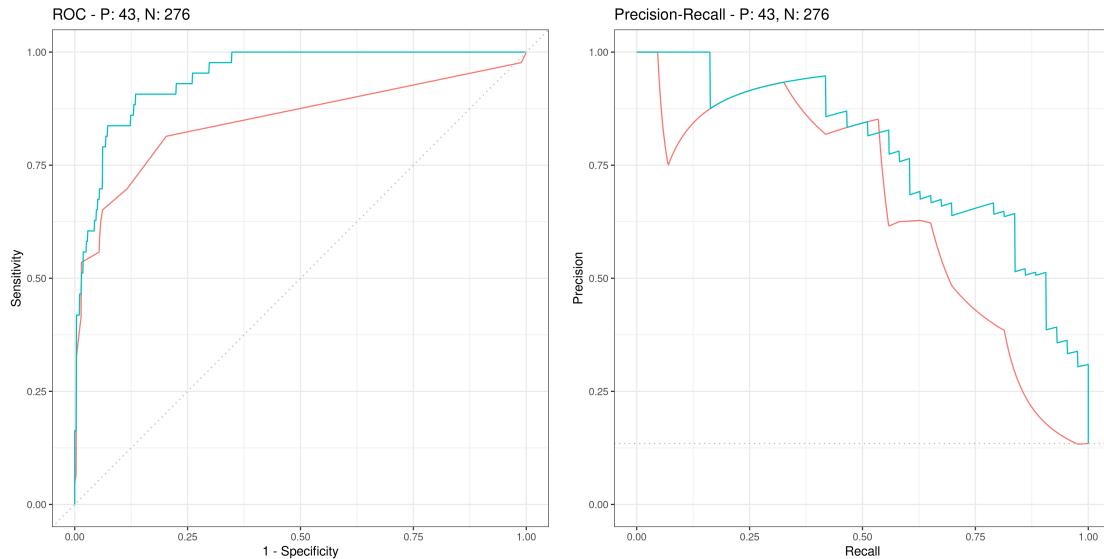


Figura 6.10: Grafici ROC e PRC per i due modelli proposti, (cart: rosso, radiale: blu)

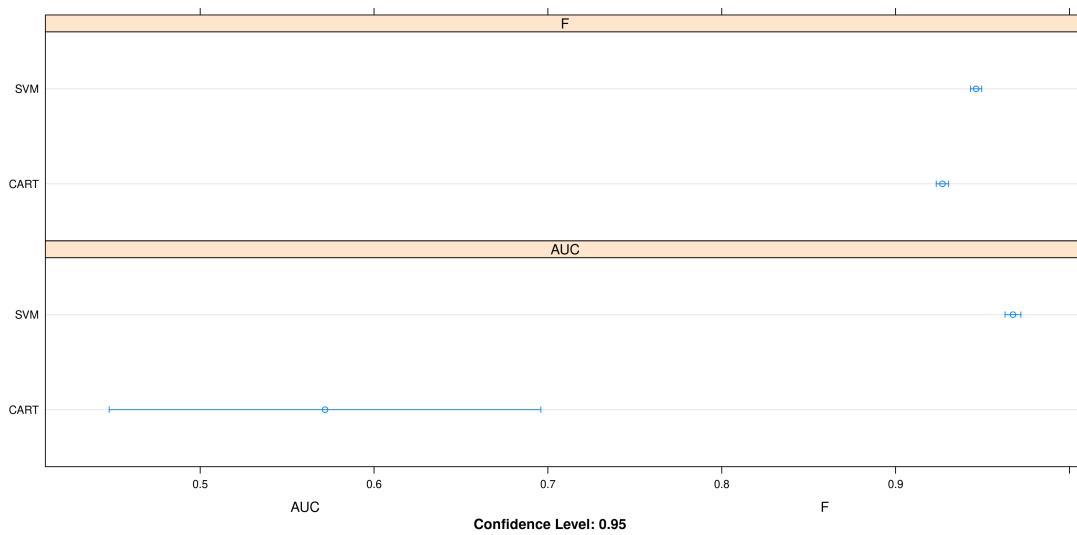


Figura 6.11: Intervalli di confidenza dei modelli proposti per AUC e F1, (svmRadial: svm radiale, rpart2: cart)

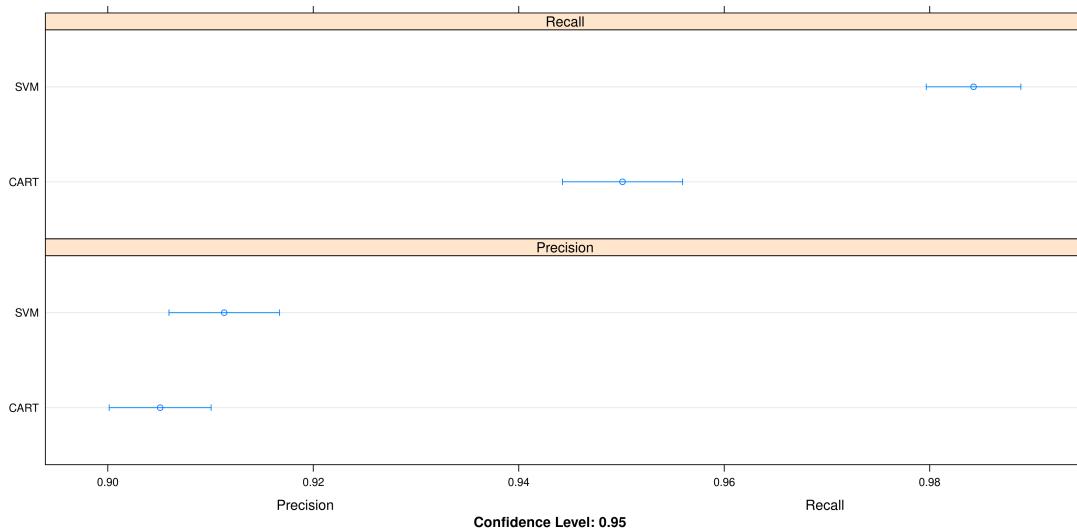


Figura 6.12: Intervalli di confidenza dei modelli proposti per Precision e Recall, (svmRadial: svm radiale, rpart2: cart)

Dagli intervalli di confidenza al 95% sono visibili leggere differenze in Precision e F1, si può notare una differenza maggiore nella Recall. CART ha un CI molto grande rispetto all'AUC.

Infine osservando i tempi di addestramento vediamo come CART prevede un tempo molto basso di 8.649s, al contrario l'SVM radiale prevede un costo più oneroso di 1827.793s. Dovendo scegliere un modello tra i due proposti l'SVM risulta migliore, nonostante abbia performance simili a CART ed un costo più elevato, poiché CART ha un intervallo di confidenza al 95% troppo grande rispetto all'AUC.

Capitolo 7

Conclusioni

In questa trattazione si è visto un problema di classificazione per la qualità del vino. Si è visto che con 10 classi il problema era difficile per la mancanza di dati per alcune classi. Si è passato quindi ad un problema binario con vini di buona qualità e vini di cattiva qualità, inoltre si è preferito analizzare solo una tipologia di vino, limitandosi al vino rosso.

Durante la fase di analisi dei dati si è visto come i dati siano sbilanciati ed alcuni attributi presentassero distribuzioni fortemente asimmetriche.

Sono stati identificati gli outliers e si è notato che la loro numerosità costituisce solamente il 2.88% del training set. E' stato deciso inizialmente di non rimuoverli ed aspettare la fase di esperimenti per prendere una decisione in seguito ai risultati ottenuti.

Analizzando le correlazioni abbiamo notato bassi valori tra i vari attributi e la qualità, il che comporta una difficoltà nel distinguere le due classi. Inoltre con la rimozione degli outliers aumentano leggermente i valori di correlazione, anche se rimangono abbastanza medio bassi.

E' stata effettuata una PCA e come previsto dai valori bassi di correlazione non ci sono stati grandi miglioramenti. Con il 95% di varianza spiegata vengono tenuti 9 degli 11 attributi.

Conclusioni

I modelli che sono stati scelti per questa trattazione sono CART e SVM, per quest'ultimo sono stati analizzati diversi kernel. Questi modelli di solito sono abbastanza robusti agli outliers e sono addestrabili anche con relativamente pochi dati. Dato che l'SVM richiede che venga effettuato un pre processing sui dati, è stato eseguito una standardizzazione con z-score.

Addestrare questi modelli ha previsto alcune accortezze a causa dello sbilanciamento dei dati. E' stata usata una 5-fold cross validation stratificata con 5 ripetizioni per ottenere risultati più attendibili. La metrica usata per la scelta del modello migliore è l'AUC della PRC.

Dagli esperimenti è risultato che il kernel migliore per SVM è quello radiale. Questo modello è stato confrontato con CART al variare del pre processing. I due modelli migliori identificati sono SVM radiale con Standardizzazione + PCA e CART con Standardizzazione, entrambi mantenendo gli outliers.

I modelli proposti presentano performance molto simili con alti valori di Precision e bassi di Recall. Le uniche differenze sono nei costi di addestramento e gli intervalli di confidenza sull'AUC della PRC, motivo per il quale, tra i due, il migliore risulta essere l'SVM.

I risultati ottenuti non sono ottimi, comunque comprensibili, poichè i dati sono sbilanciati e le feature poco discriminanti. Possibili sviluppi futuri, oltre la raccolta di nuovi dati, possono essere gestire lo sbilanciamento con tecniche di undersampling o oversampling esempio SMOTE [2], altri modelli più sofisticati possono essere utilizzati come le Random Forest [1]. Inoltre si potrebbe estendere il problema a più classi o considerare anche il vino bianco.

Bibliografia

- [1] Gérard Biau e Erwan Scornet. «A random forest guided tour». In: *Test* 25.2 (2016), pp. 197–227.
- [2] Nitesh V Chawla et al. «SMOTE: synthetic minority over-sampling technique». In: *Journal of artificial intelligence research* 16 (2002), pp. 321–357.
- [3] Wikimedia Commons. *Interquartile Range*. 2021. URL: https://upload.wikimedia.org/wikipedia/commons/1/1a/Boxplot_vs_PDF.svg.
- [4] Wikimedia Commons. *Vino*. URL: <https://it.wikipedia.org/wiki/Vino>.
- [5] Corinna Cortes e Vladimir Vapnik. «Support-vector networks». In: *Machine learning* 20.3 (1995), pp. 273–297.
- [6] Paulo Cortez et al. «Modeling wine preferences by data mining from physicochemical properties». In: *Decision support systems* 47.4 (2009), pp. 547–553.
- [7] DeepAI. *kurtosis*. URL: <https://deeppai.org/machine-learning-glossary-and-terms/kurtosis>.
- [8] DeepAI. *skewness*. URL: <https://deeppai.org/machine-learning-glossary-and-terms/skewness>.
- [9] Lindsay I Smith. «A tutorial on principal components analysis». In: (2002).
- [10] Wikipedia. *Analisi delle componenti principali*. URL: https://it.wikipedia.org/wiki/Analisi_delle_componenti_principali.