Progetto Machine Learning: Red Wine Quality

829612 Magazzù Giuseppe
829685 Magazzù Gaetano
829889 Malanchini Mirco

UCI machine learning repository

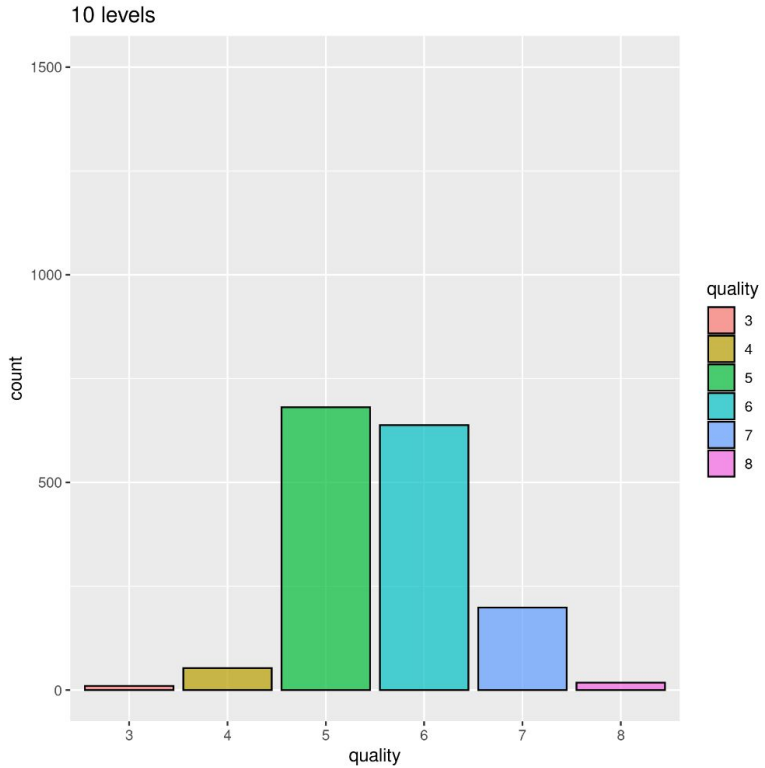
Wine Quality Data Set

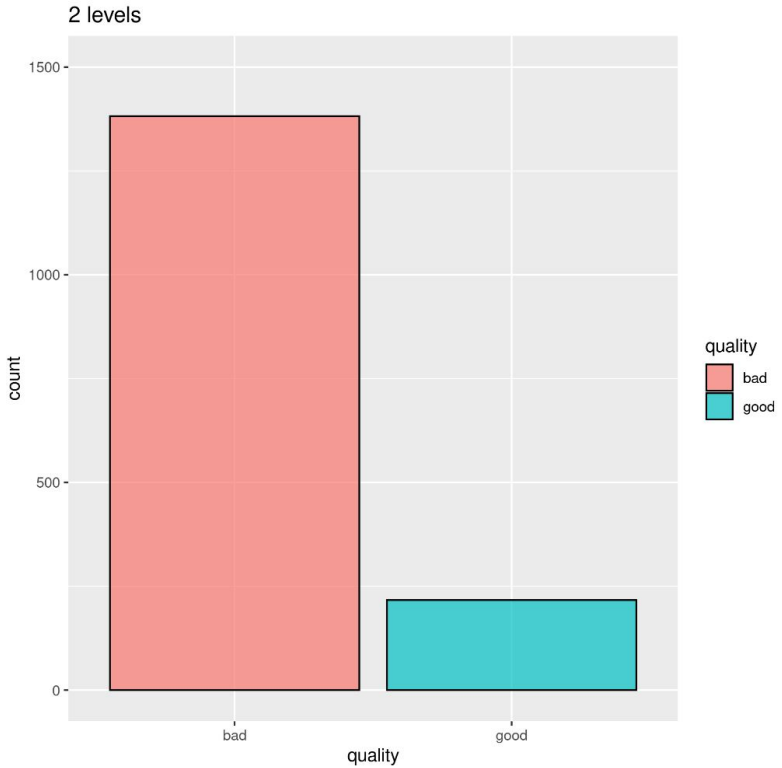
- **11 Attributi + 1 Output**
  fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, **quality**

- **Red: 1599 Osservazioni**
- **White: 4989 Osservazioni**

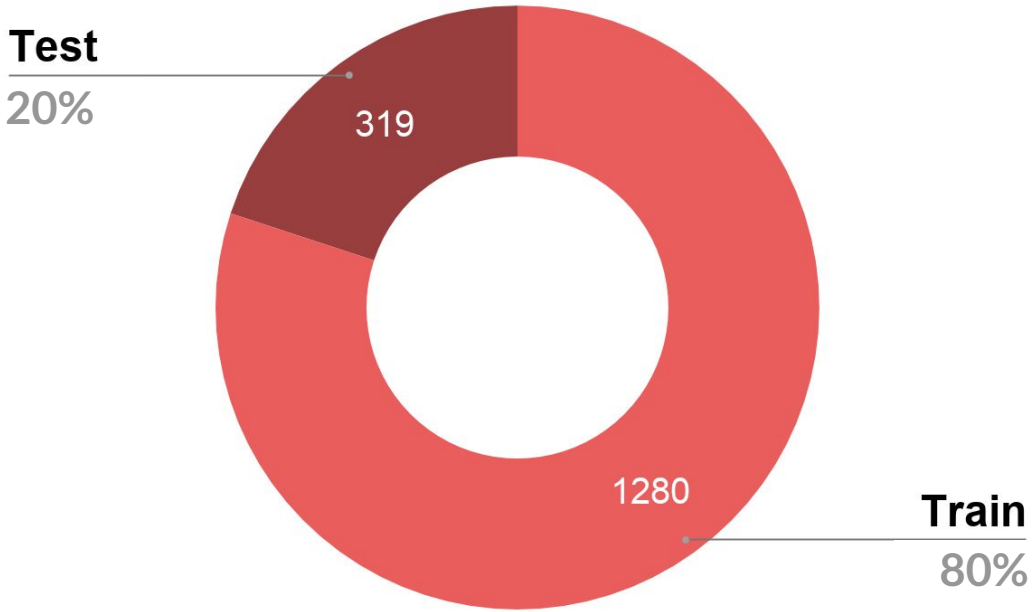
P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

Attributo Quality





Partizionamento del dataset



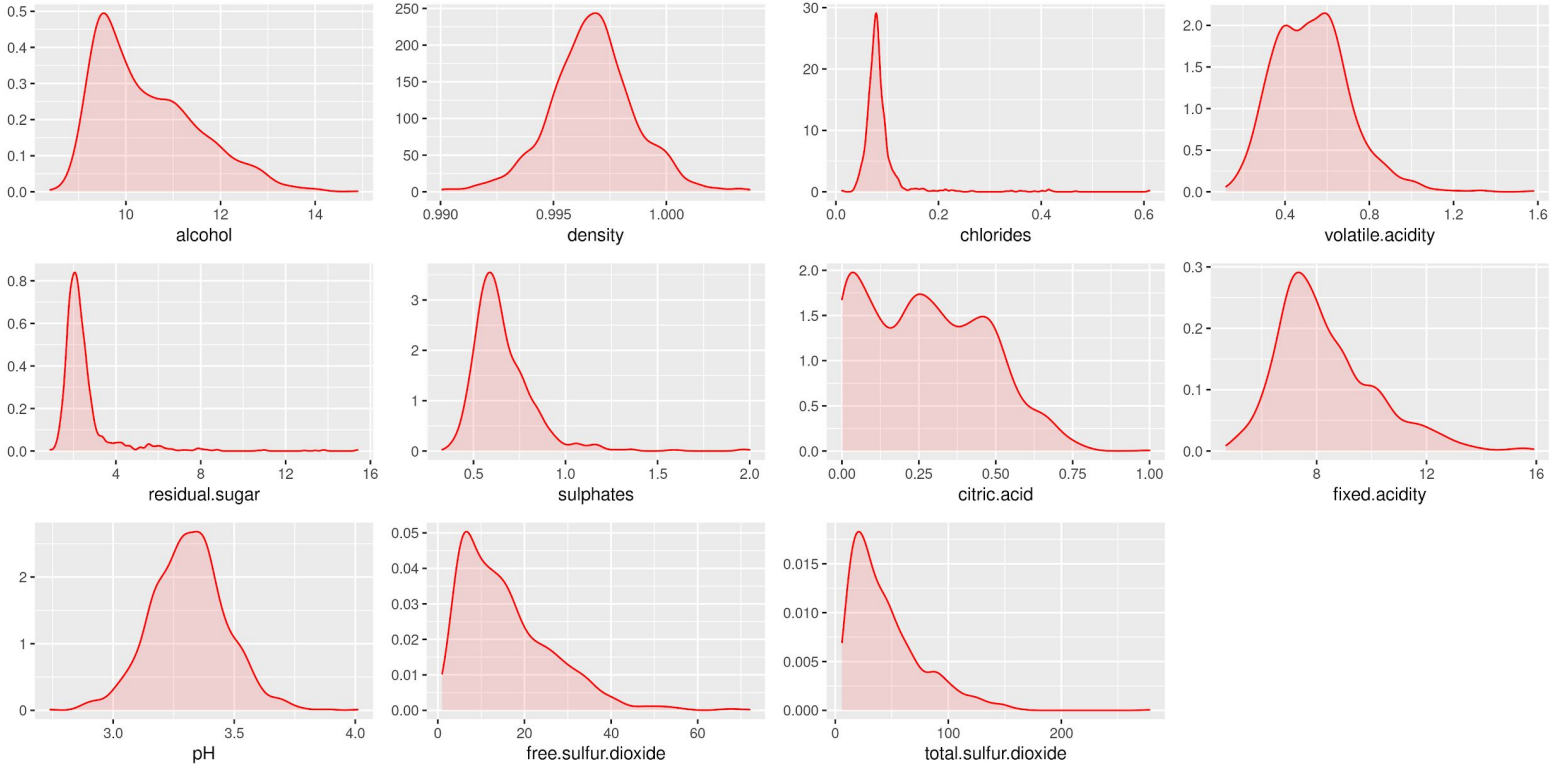
Exploratory Data Analysis

Analisi Univariata

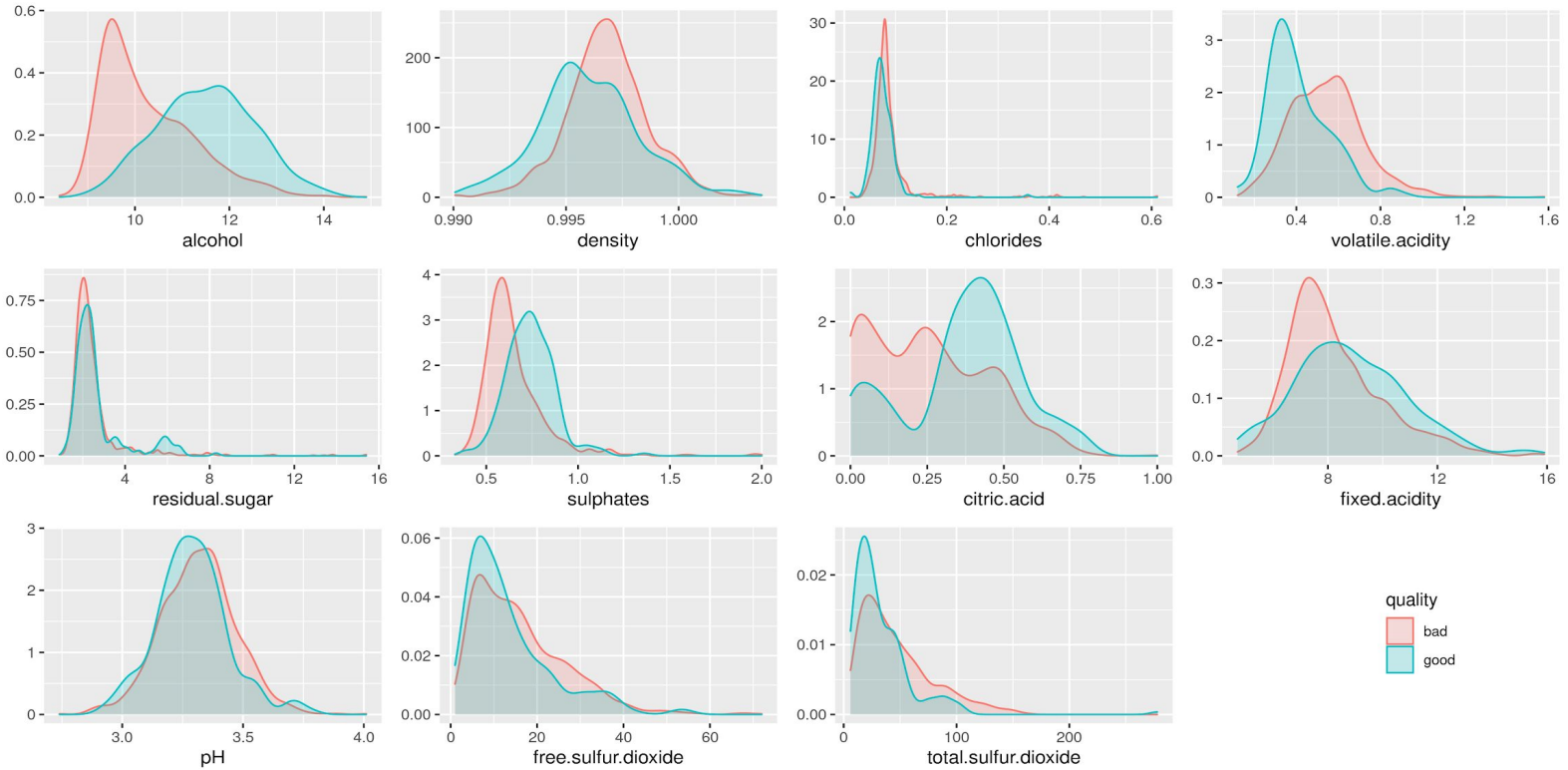
- Valori mancanti
- Statistiche descrittive
- Analisi distribuzioni
- Outliers

	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	0,02	0,01	0,02	0,01	0,05	0,09	0,26
citric.acid	0	0,33	0,12	0,32	0,00	1,66	1,35	6,80
residual.sugar	0	0,28	0,22	0,22	0,04	2,76	0,05	0,18
chlorides	0	0,05	0,02	0,04	0,01	0,35	5,12	39,50
free.sulfur.dioxide	0	1,48	0,72	1,42	0,08	12,04	0,08	0,58
total.sulfur.dioxide	0	138,36	42,60	134,00	9,00	440,00	0,42	0,78
density	0	0,07	0,00	0,07	0,07	0,04	0,05	0,49
pH	0	3,19	0,15	3,18	2,72	3,82	0,46	0,57
sulphates	0	0,03	0,01	0,03	0,02	0,05	0,04	0,09
alcohol	0	10,51	1,23	10,40	8,00	14,20	0,51	-0,67

Analisi Univariata



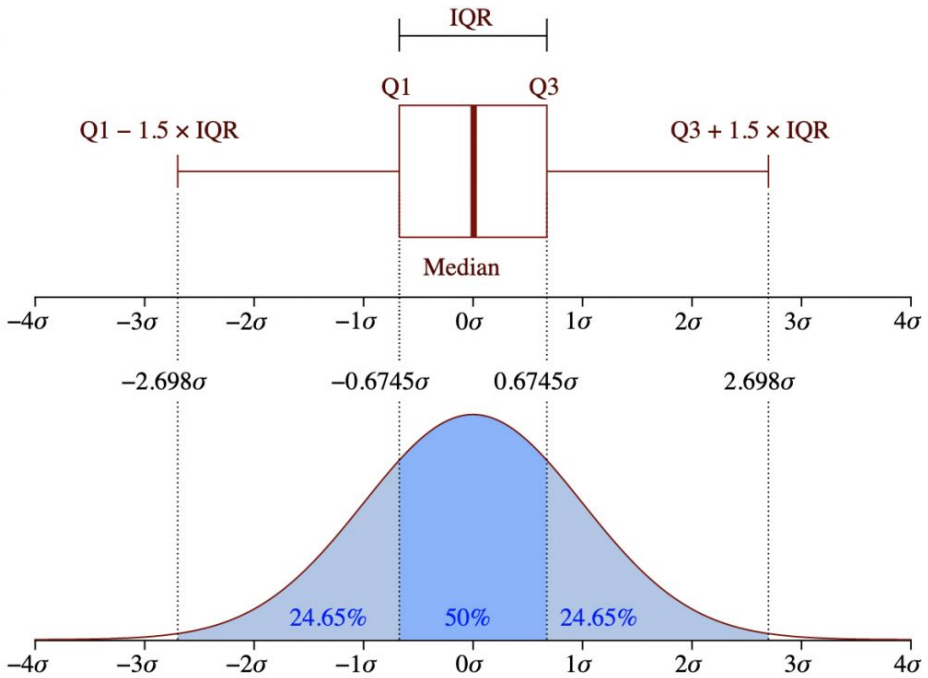
Analisi Univariata



Analisi Outliers

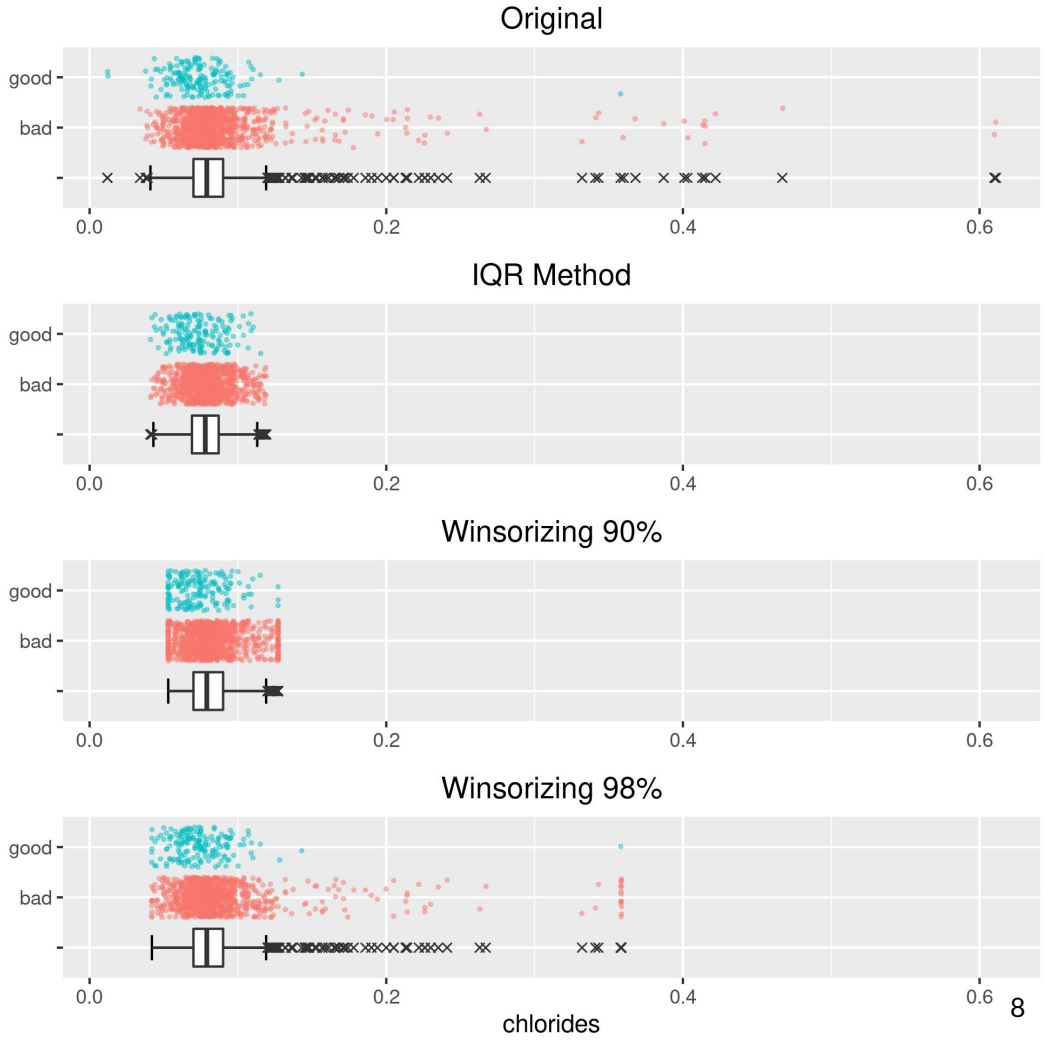
2 Metodi Statistici:

- Interquartile Range (rimozione)
- Winsorizing (capping)
  - Winsorizing 90% (0.05, 0.95)
  - Winsorizing 98% (0.01, 0.99)



Analisi Outliers

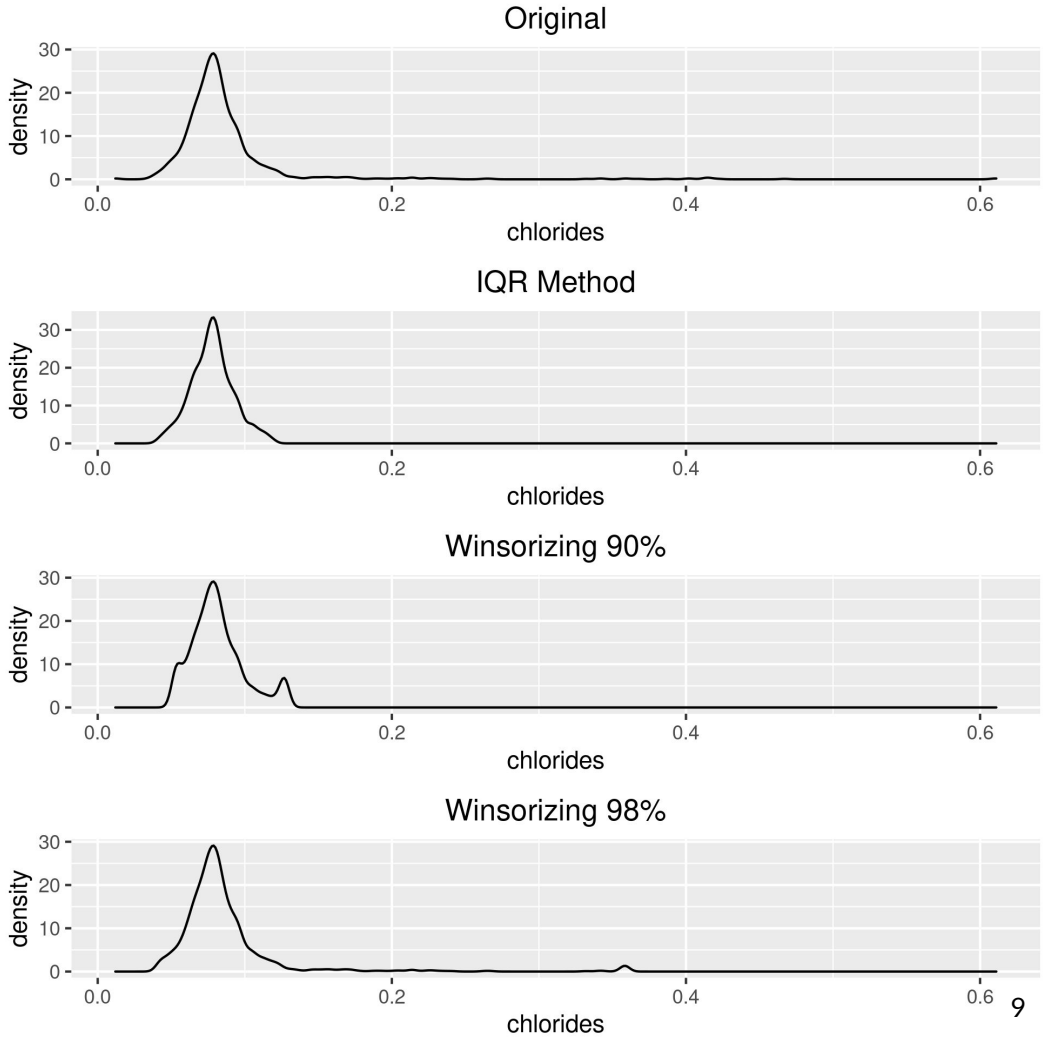
- Tutte le variabili hanno outliers
- Pochi outliers per la classe good
- Skewness positiva => Molti outliers con valori alti
- Sulphates, residual.sugar, total.sulfur.dioxide e chlorides sono le variabili con più outliers



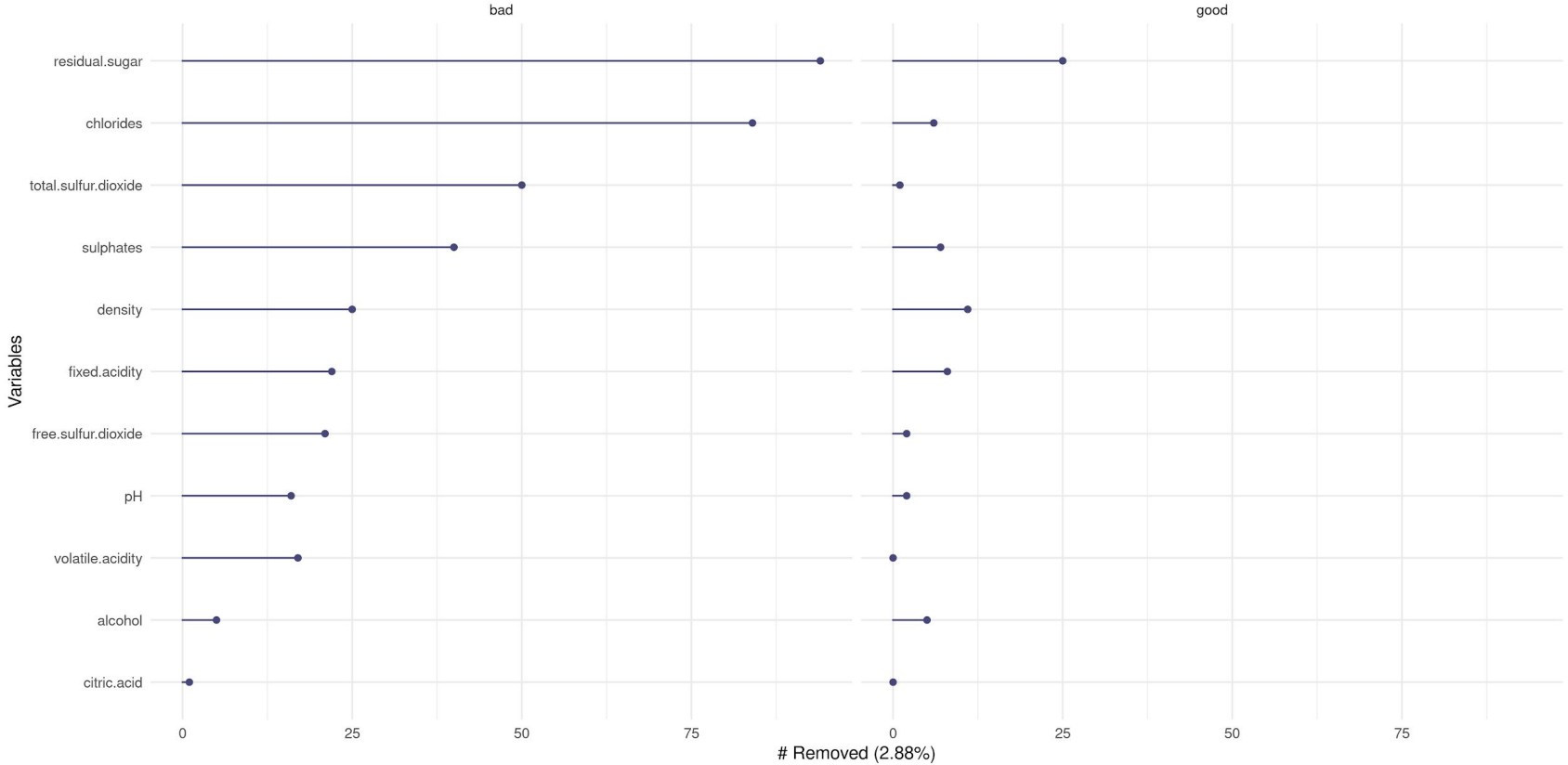
Analisi Outliers

- Winsorizing deforma l'istogramma
- Winsorizing 98% tiene alcuni outliers
- Con IQR, residual.sugar, chlorides e sulphates assumono una distribuzione quasi simmetrica

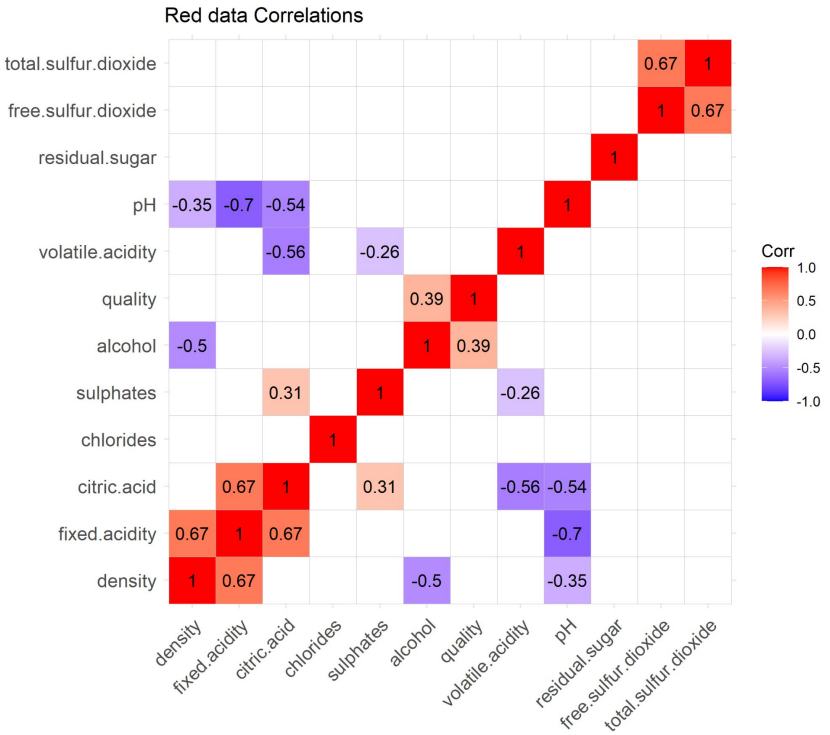
Metodo Scelto: IQR

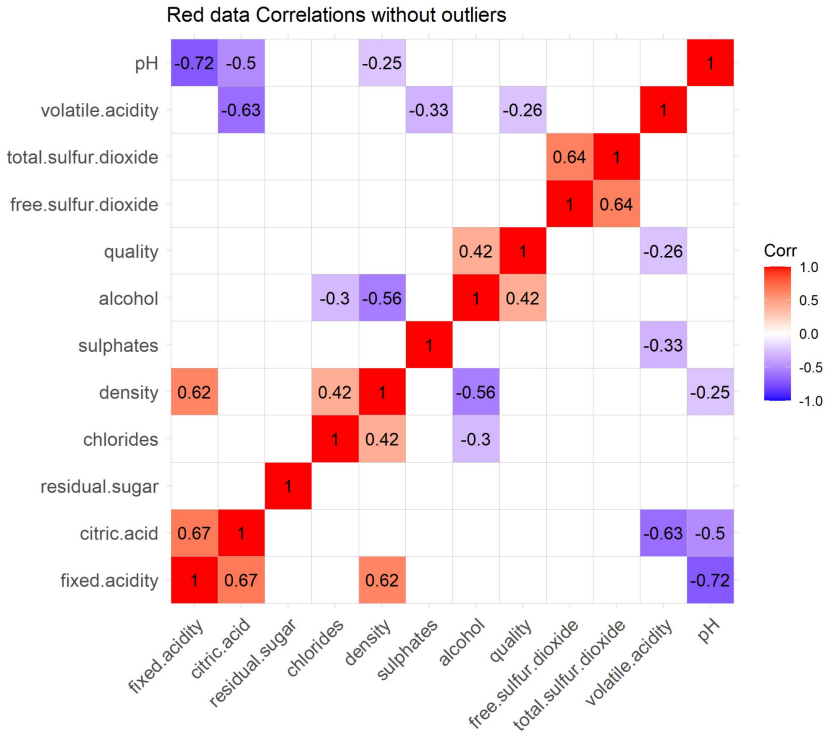


Analisi Outliers



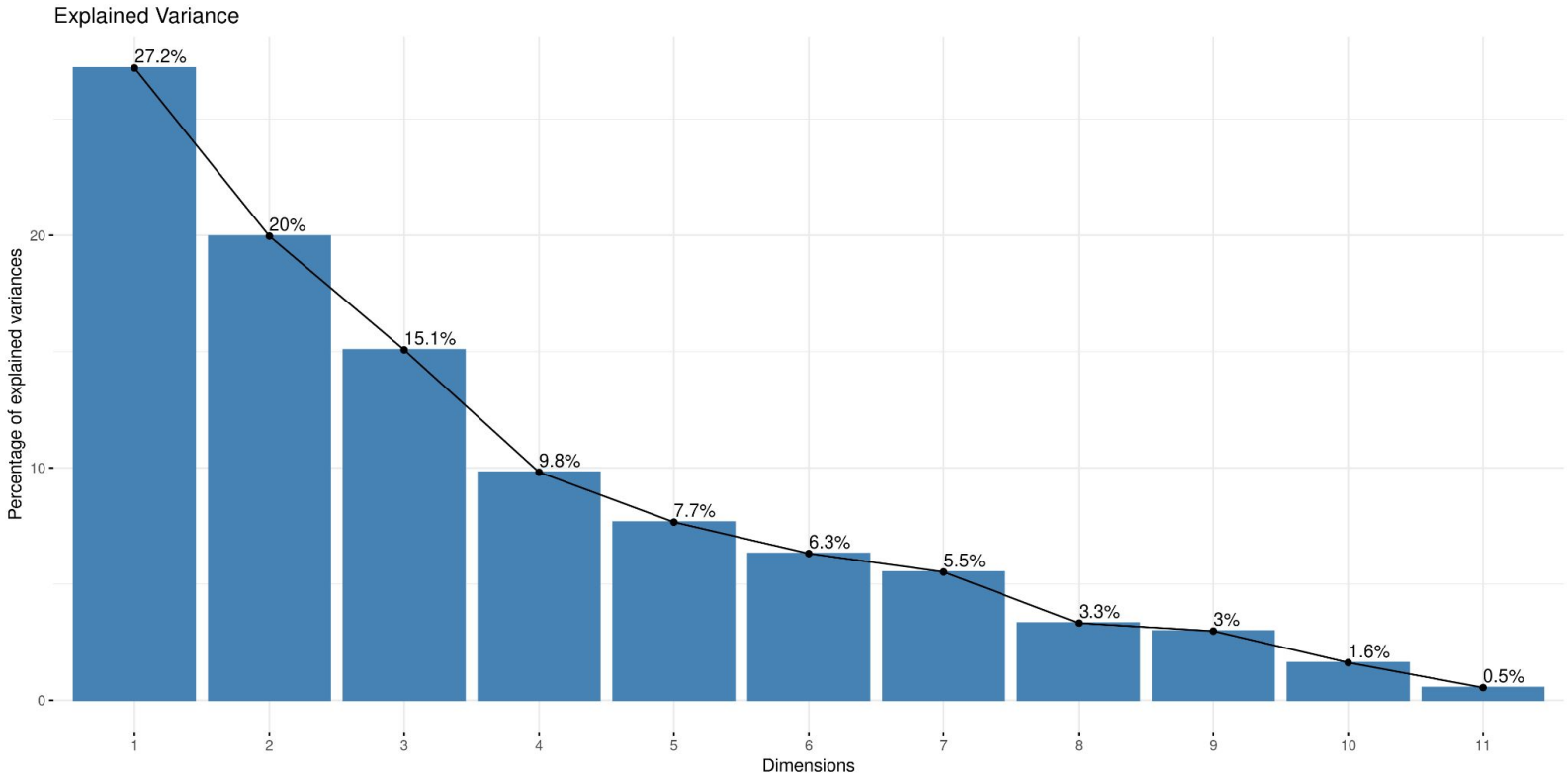
Analisi Multivariata





Analisi Multivariata

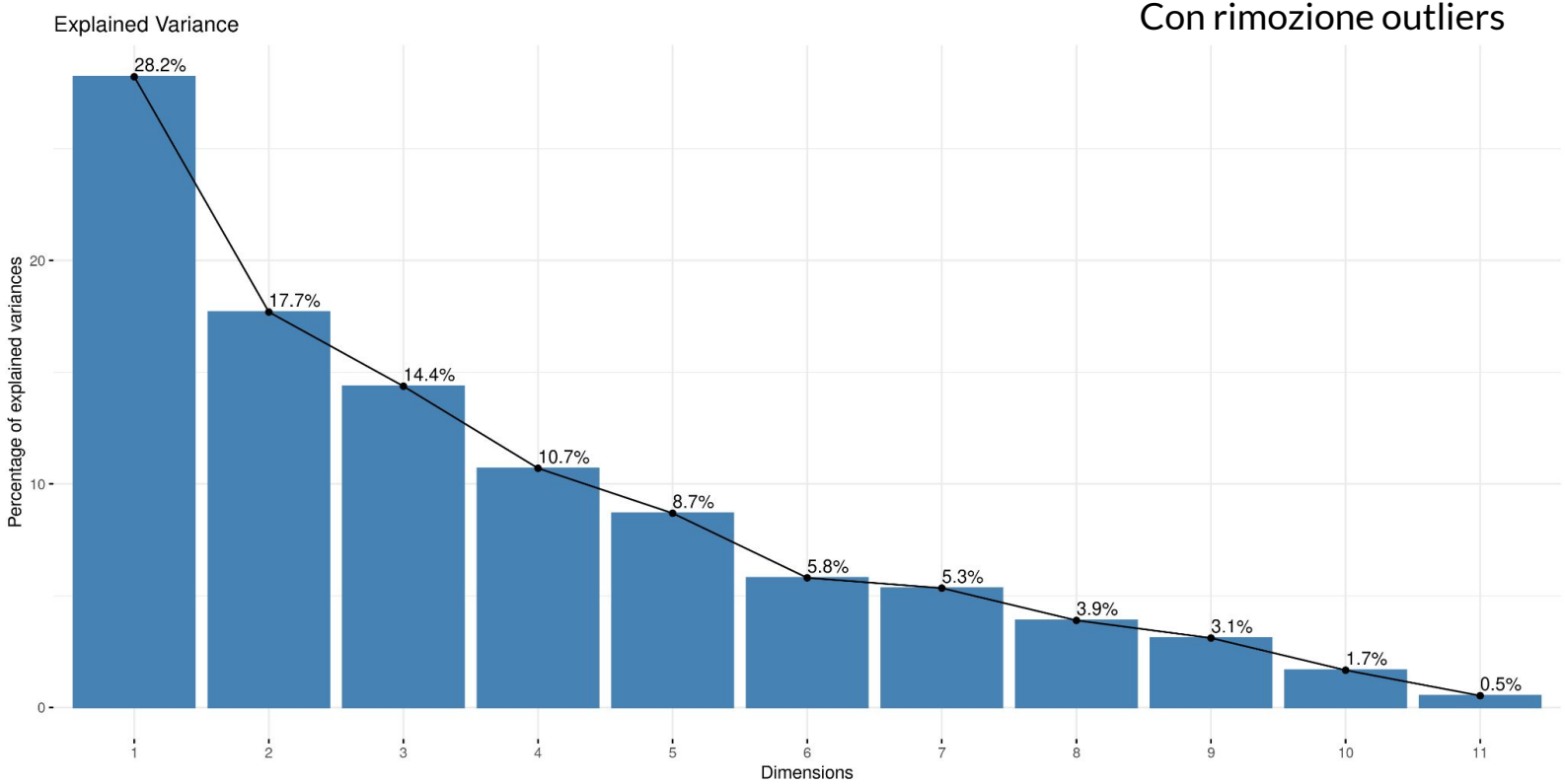
Threshold 95%, 8 componenti



Analisi Multivariata

Threshold 95%, 8 componenti

Con rimozione outliers



Pre Processing

Pre Processing

Metodo 1

\+ Standardizzazione (z-score)

Metodo 2

\+ Standardizzazione (z-score)

\+ PCA

Rimozione Outliers

Pre Processing

Datasets

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

Modelli e Addestramento

Modelli e Addestramento

- 5-fold cross validation stratificata
- 5 ripetizioni
- Ottimizzazione AUC-PRC per la scelta del modello migliore
- Tuning con Grid Search

SVM

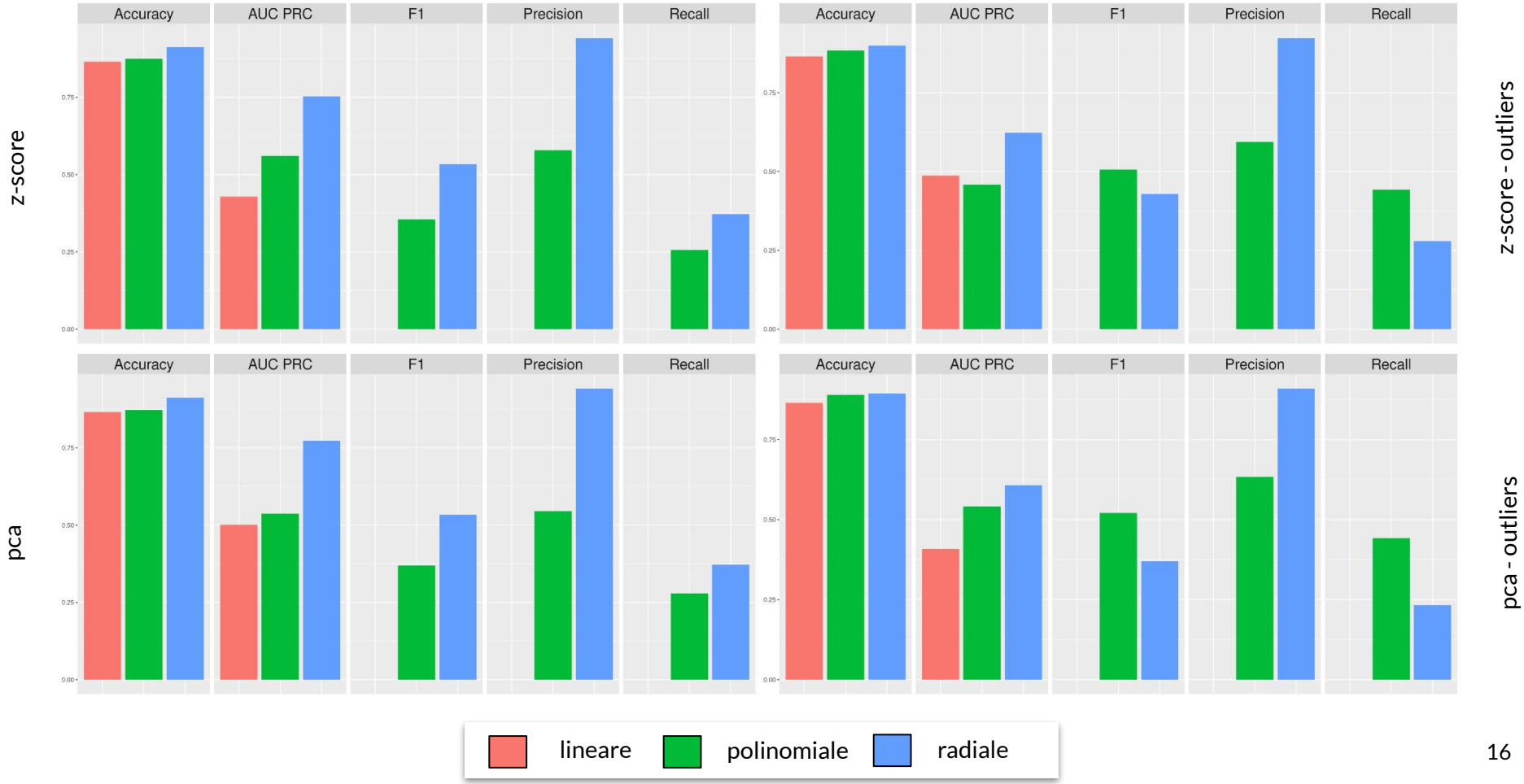
- Nella versione soft margin permette una certa tolleranza agli outliers
- Richiede un pre processing
- Richiede meno dati di una rete neurale
- Al contrario di altri modelli come Naive Bayes e CART ha un costo computazionale alto
- Kernel (Lineare, Polinomiale, Radiale)

CART

- Poco soggetto agli outliers e ai valori mancanti
- Non richiede pre processing
- Poco costoso computazionalmente rispetto a SVM e Reti Neurali
- Per alberi profondi richiede utilizzo di tecniche di pruning per evitare problematiche di overfitting

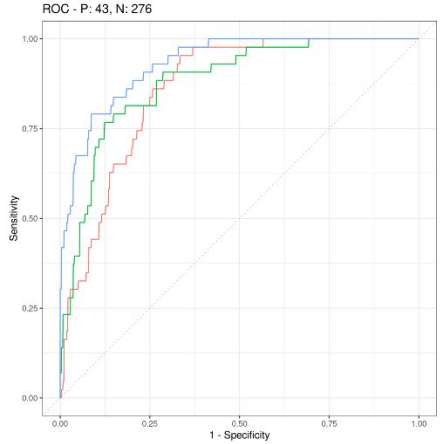
Risultati

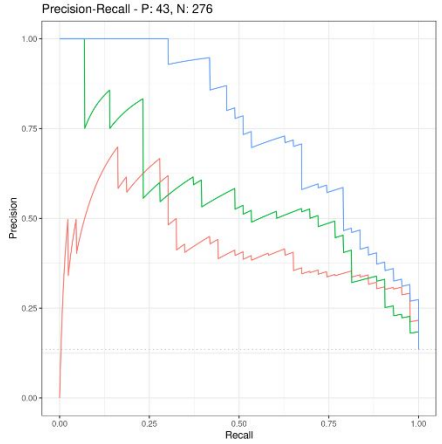
Confronto Kernel SVM

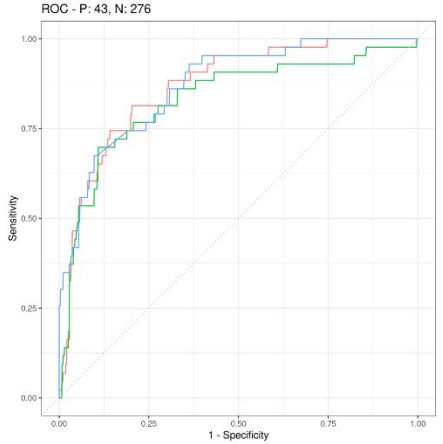


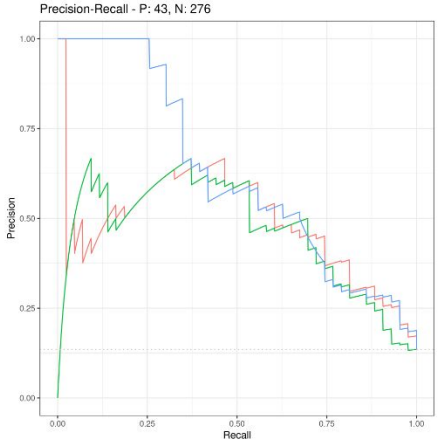
Confronto Kernel SVM

z-score



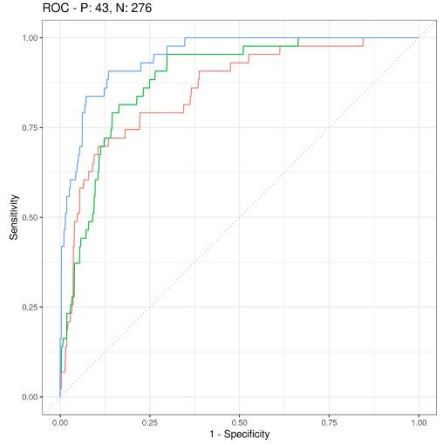


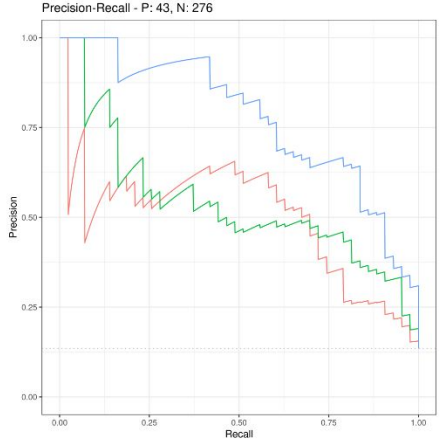


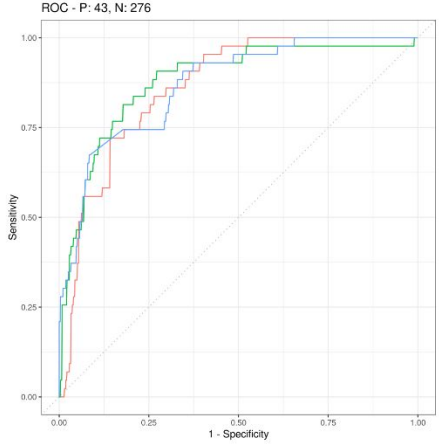


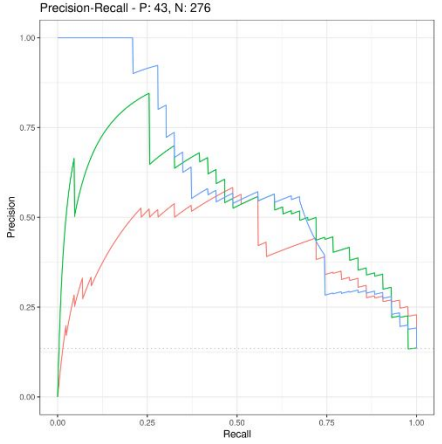
z-score - outliers

pca





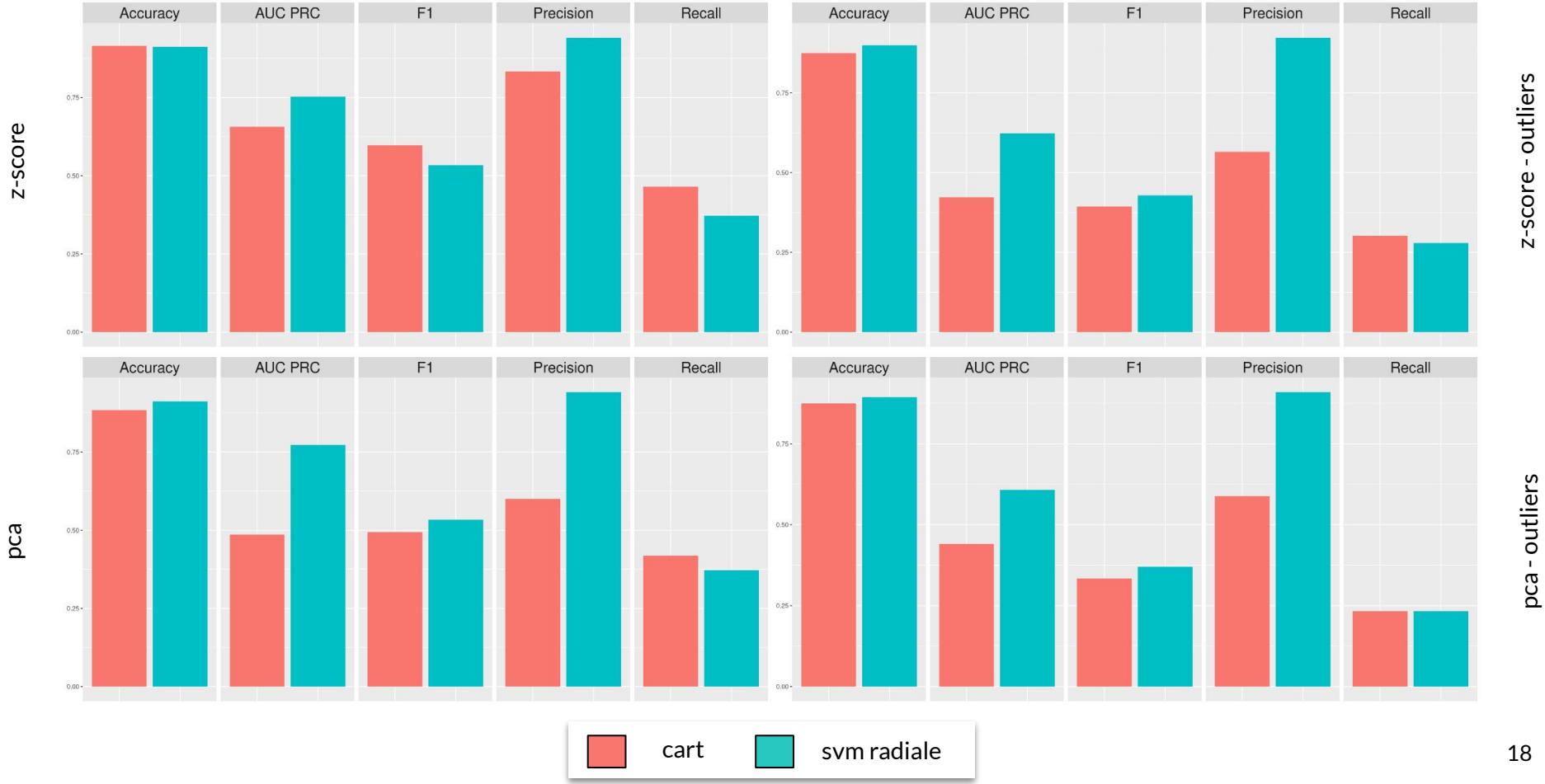




pca - outliers

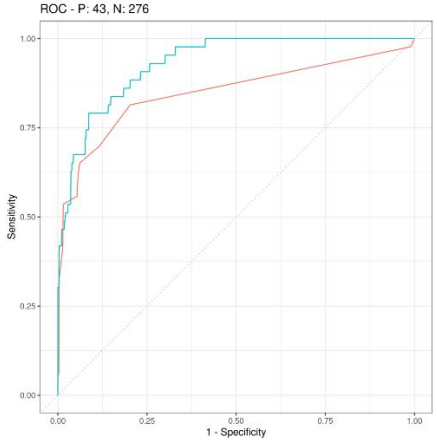
lineare polinomiale radiale

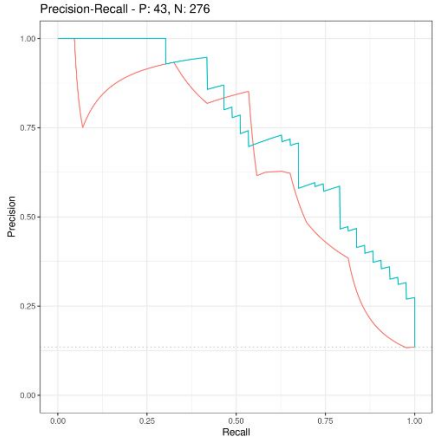
Confronto SVM radiale e CART

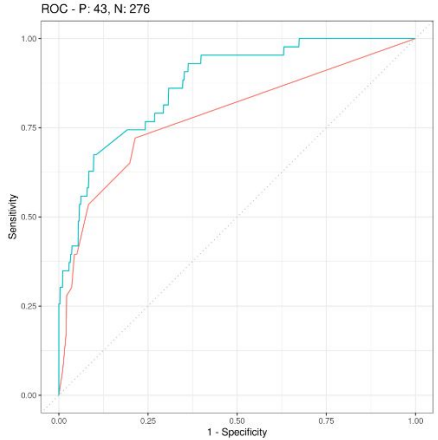


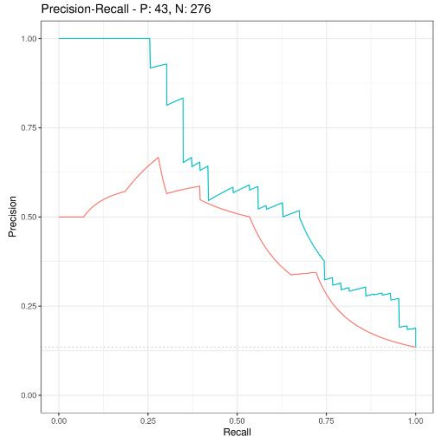
Confronto SVM radiale e CART

z-score



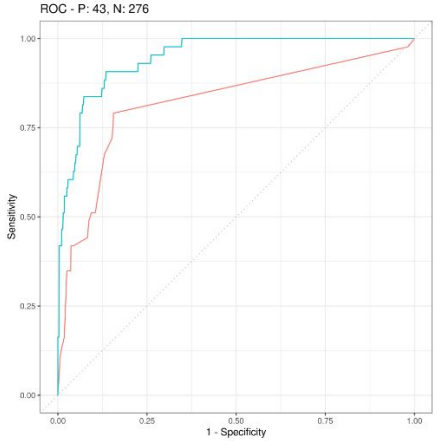


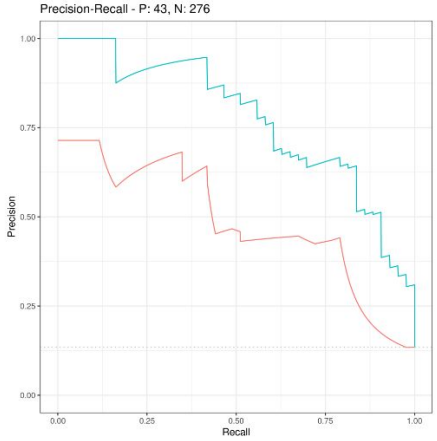


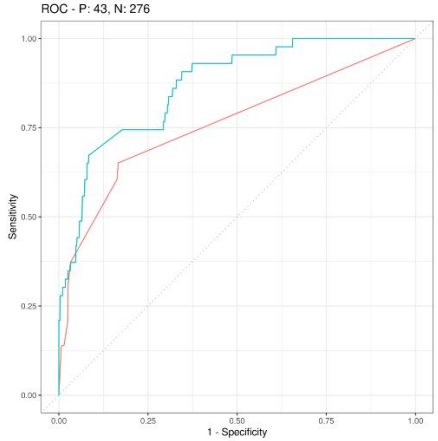


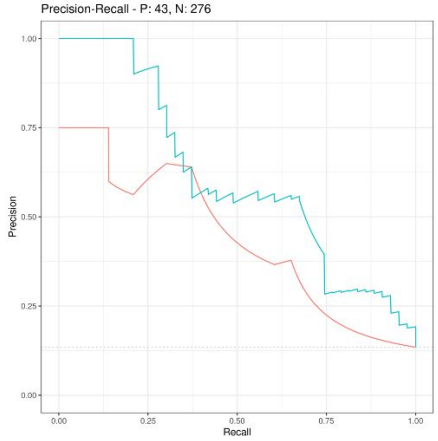
z-score - outliers

pca









pca - outliers

cart

svm radiale

Confronto Modelli

Matrici Di Confusione

Test Set - 319 istanze (negative: 276, positive: 43)

Accuracy: 0.91

Precision: **0.94**

Recall: 0.37

F1: 0.50

AUC PRC: **0.77**

SVM Radiale (PCA)

Accuracy: **0.92**

Precision: 0.88

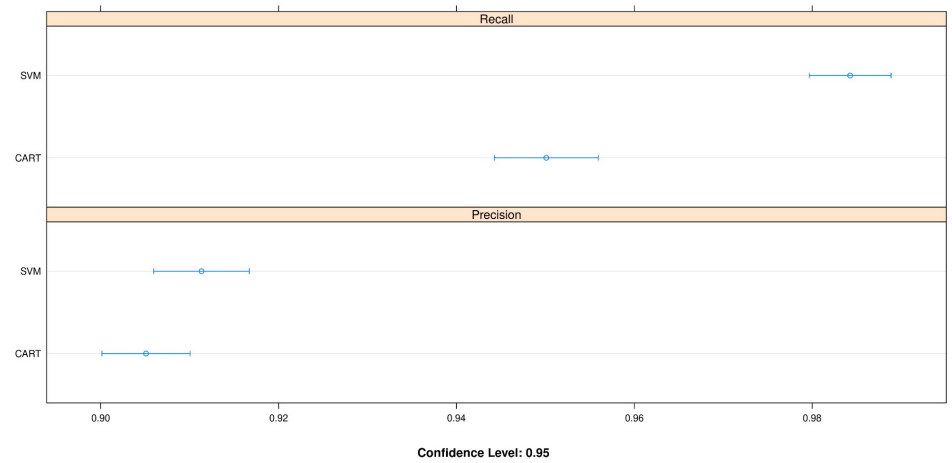
Recall: **0.47**

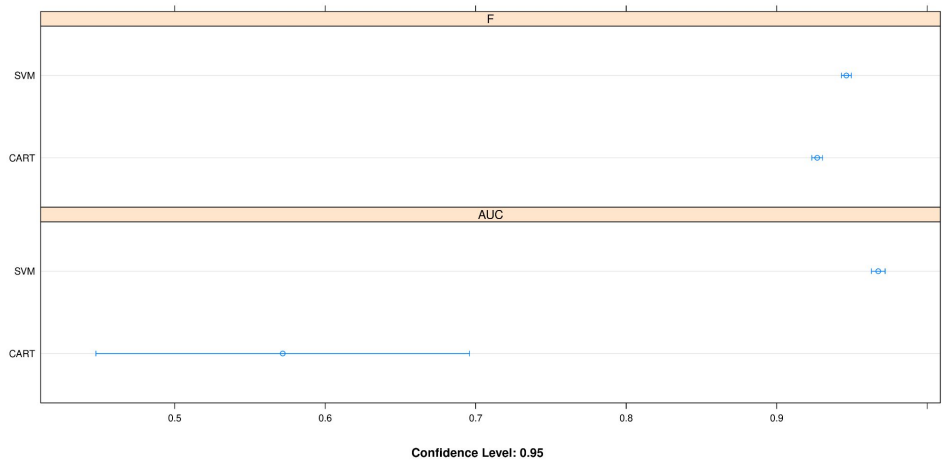
F1: **0.56**

AUC PRC: 0.75

CART (z-score)

Intervalli Di Confidenza 95%



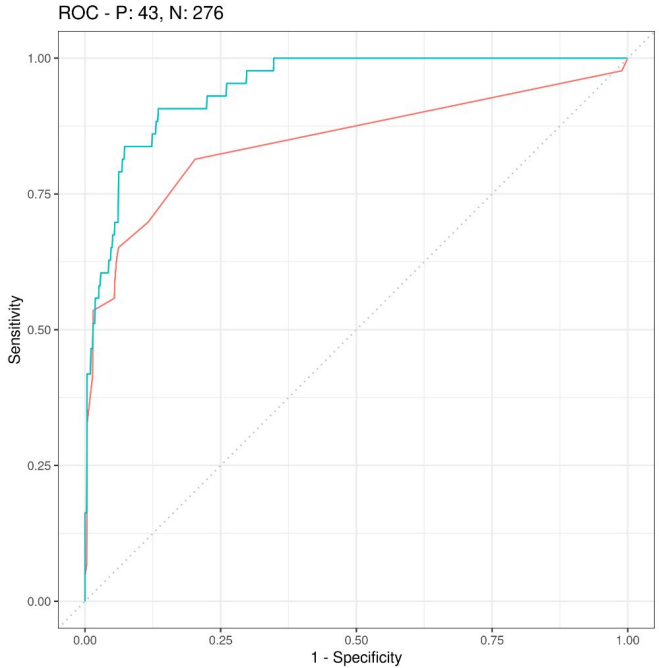


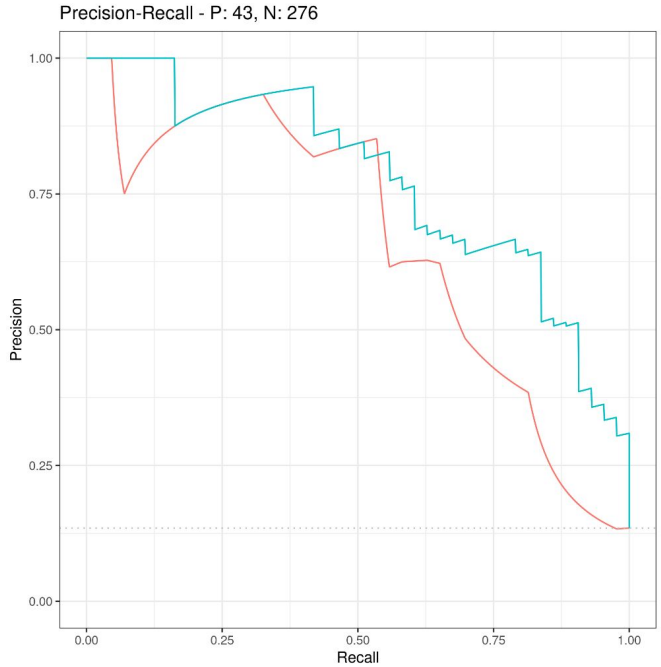
Curva ROC e PRC

	SVM	CART
ROC (AUC)	0,95	0,93
PRC (AUC)	0,77	0,75
Training Time (+ Tuning)	1827s	8,06s

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

$$\text{1 - Specificity} = \frac{FP}{FP+TN}$$





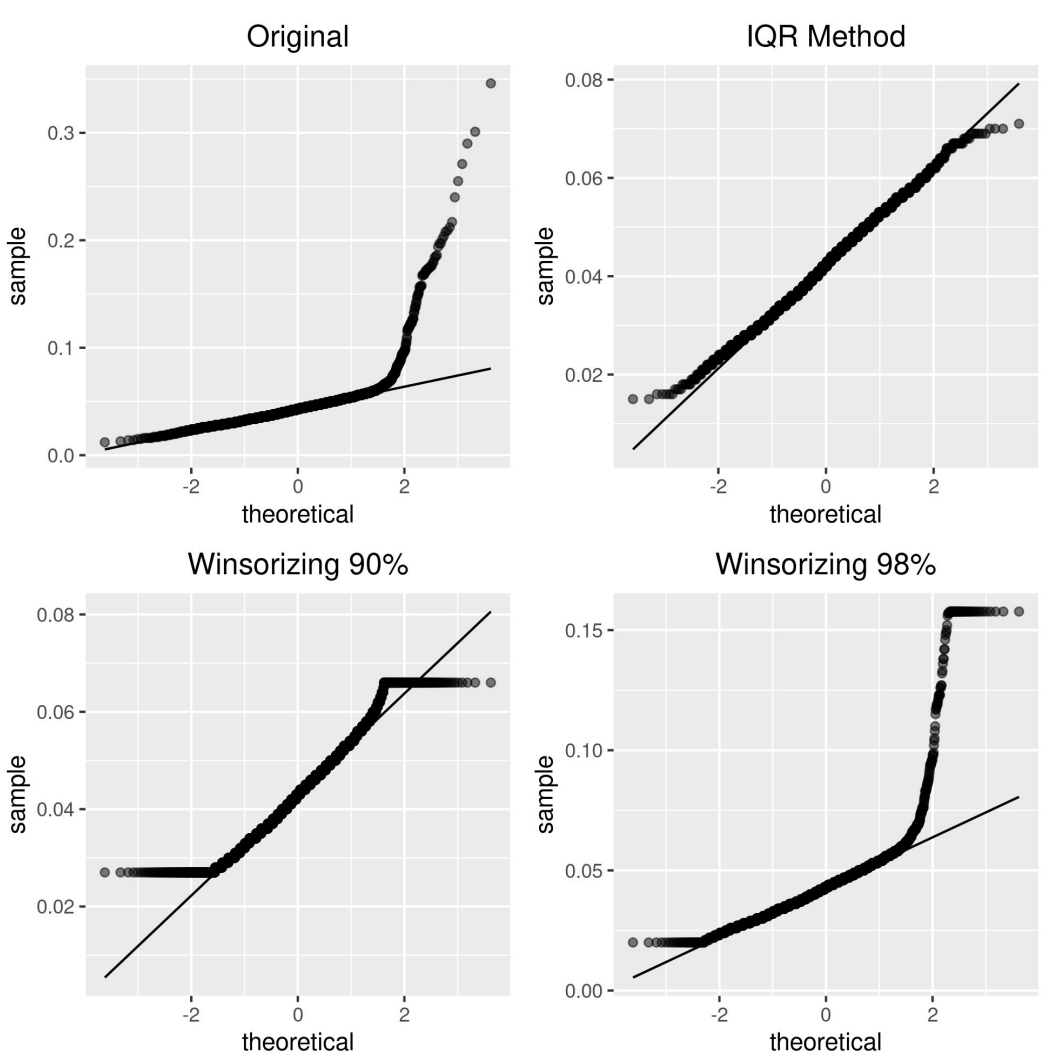
cart (z-score) svm (pca)

Conclusioni

1. Features poco discriminanti
2. Rimuovere gli outliers non migliora i modelli
3. In generale la PCA non migliora i risultati
4. Dati fortemente sbilanciati, dati skewed
5. I modelli presentati hanno performance simili, ma l'SVM ha un costo maggiore di CART, tuttavia gli intervalli di confidenza dell'SVM sono migliori
6. Recall bassa per la classe good su entrambi i modelli (FN alto)
7. Possibili migliorie usare tecniche di oversampling o undersampling(es. SMOTE) , oppure modelli più sofisticati (es.random forest), inoltre si potrebbe estendere il problema a più classi o considerare anche il vino bianco

Extra

Analisi Outliers: Q-Q plot



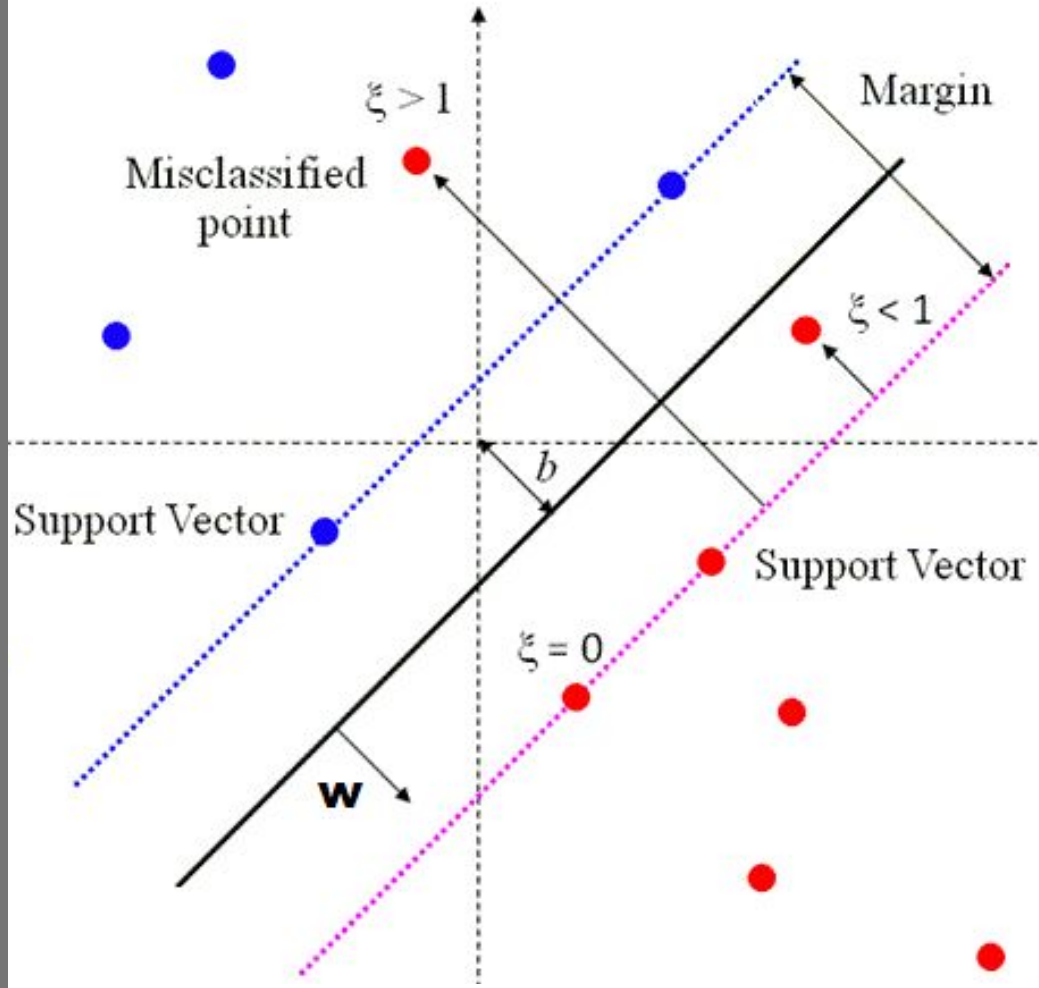
SVM

$$min \frac{1}{2}||w||^2 + C\sum_i \xi_i$$
$$y_i(w^T x_i + b) \geq 1 - \xi_i$$
$$\xi_i \geq 0$$

C: più è grande più diminuisce il margine, permette miss-classification

Gamma: controlla la forma dell'iperpiano

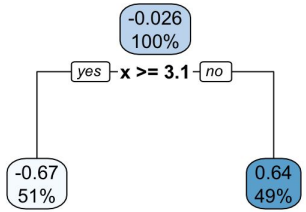
$$K(\mathbf{x}, \mathbf{x}') = \exp(-\gamma\|\mathbf{x} - \mathbf{x}'\|^2)$$
$$K(\mathbf{x}, \mathbf{x}') = (scale\langle \mathbf{x}, \mathbf{x}' \rangle + offset)^{degree}$$

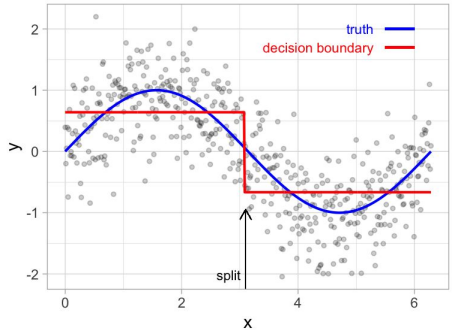


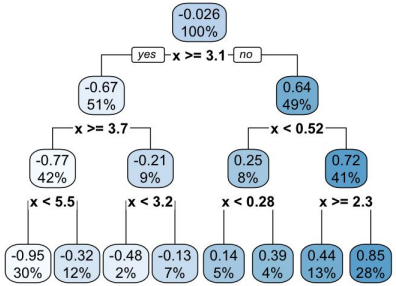
https://tomaszkacmajor.pl/index.php/2016/04/24/svm-model-selection/

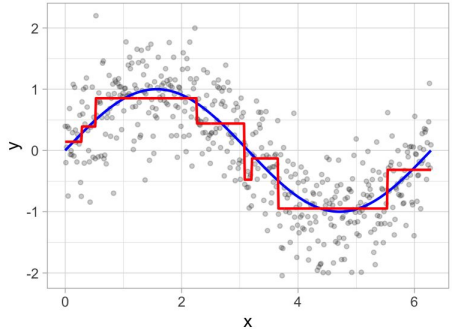
Decision Tree

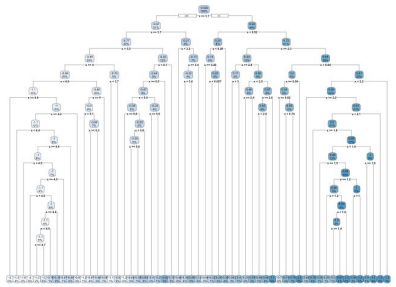
Max Depth: Tramite la scelta di questo parametro è possibile ottenere un albero meno profondo e ridurre il rischio di overfitting

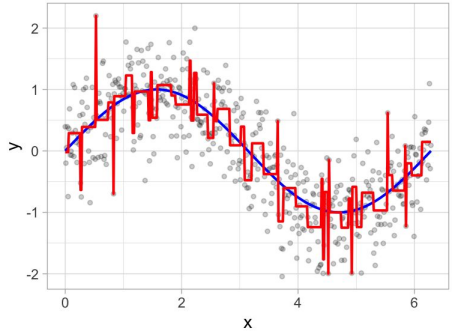












https://bradleyboehmke.github.io/HOML/DT.html

Tabelle Risultati Esperimenti

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

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