Tecniche di Machine Learning per la Selezione della Terapia nei casi di Malaria Severa

Candidato: Luca Dal Zotto

Relatore: Prof. Francesco Rinaldi

Dipartimento di Matematica "Tullio Levi-Civita"
Corso di Laurea in Matematica

27 Settembre 2019

Università degli Studi di Padova

Contenuti

Alcuni dati sulla malaria

- è tra le più importanti malattie infettive al mondo per diffusione e mortalità
- circa 3,3 miliardi di persone vivono in aree endemiche
- nel 2016 i casi accertati accertati sono stati 216 milioni

Alcuni dati sulla malaria

- è tra le più importanti malattie infettive al mondo per diffusione e mortalità
- circa 3,3 miliardi di persone vivono in aree endemiche
- nel 2016 i casi accertati accertati sono stati 216 milioni

Complicazione: **malaria severa**

Possibili terapie: **orale** o **endovenosa**

Fonte: Istituto Nazionale Malattie Infettive Lazzaro Spallanzani-IRCCS-Roma

Dati: 259 pazienti, di cui 119 casi di malaria severa

Introduzione

Presentazione del problema

Fonte: Istituto Nazionale Malattie Infettive Lazzaro Spallanzani-IRCCS-Roma

Dati: 259 pazienti, di cui 119 casi di malaria severa

Criteri ufficiali WHO	**Valori ematici all'atto del ricovero**	**Altro**
cerebral malaria/coma	PLT	età
convulsions	Hb	sesso
acute renal failure	creat	comorbidità
respiratory failure	bil	provenienza
hypoglycaemia	AST	zona in cui si è contratta l'infezione
shock	ALT	
spontaneous bleeding	Na	pregressa malaria
acidosis	parassitemia baseline	durata permanenza nel paese endemico
jaundice		
liver function test >3 time normal range		ritardo diagnosi
		ritardo accesso cure
anemia		chemioprofilassi
hyperparasitemia		

Tabella: Elenco completo delle feature considerate per ogni paziente.

Strumento usato per l'analisi: libreria Python Scikit-Learn

Strumento usato per l'analisi: libreria Python Scikit-Learn

- organizzazione dei dati in **feature matrix** e **target vector**

Strumento usato per l'analisi: libreria Python Scikit-Learn

- organizzazione dei dati in **feature matrix** e **target vector**
- gestione dei **missing values**

Strumento usato per l'analisi: libreria Python Scikit-Learn

- organizzazione dei dati in **feature matrix** e **target vector**
- gestione dei **missing values**
- codifica **one-hot** per le feature categoriche

Strumento usato per l'analisi: libreria Python Scikit-Learn

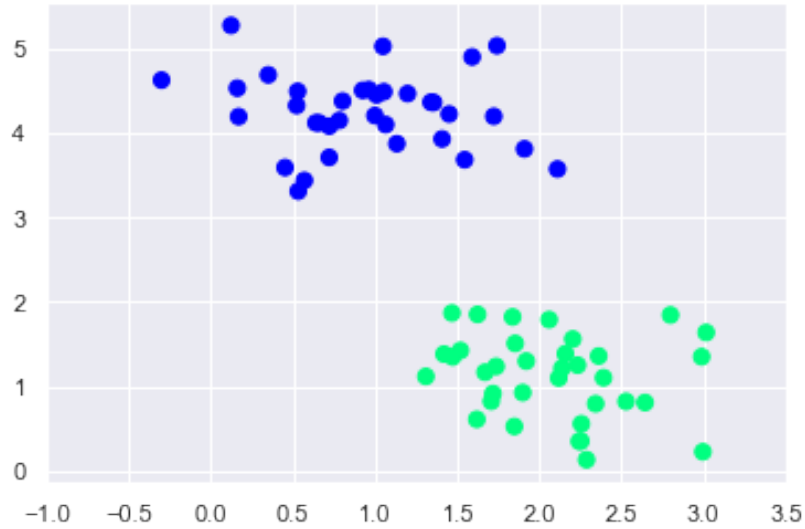
- organizzazione dei dati in **feature matrix** e **target vector**
- gestione dei **missing values**
- codifica **one-hot** per le feature categoriche
- **normalizzazione** dei dati

Apprendimento supervisionato

Attraverso una fase di addestramento in cui si tengono in considerazione gli output corretti, viene approssimata la funzione che mappa ogni paziente nella sua terapia.

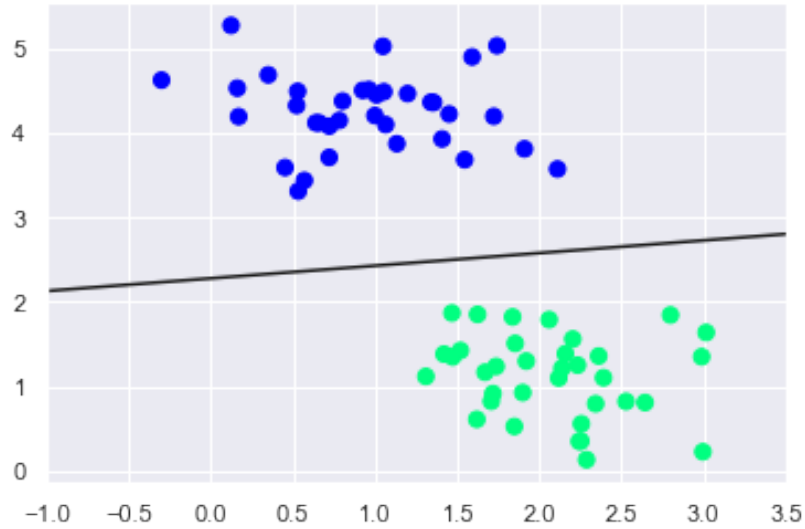
Apprendimento supervisionato

Attraverso una fase di addestramento in cui si tengono in considerazione gli output corretti, viene approssimata la funzione che mappa ogni paziente nella sua terapia.



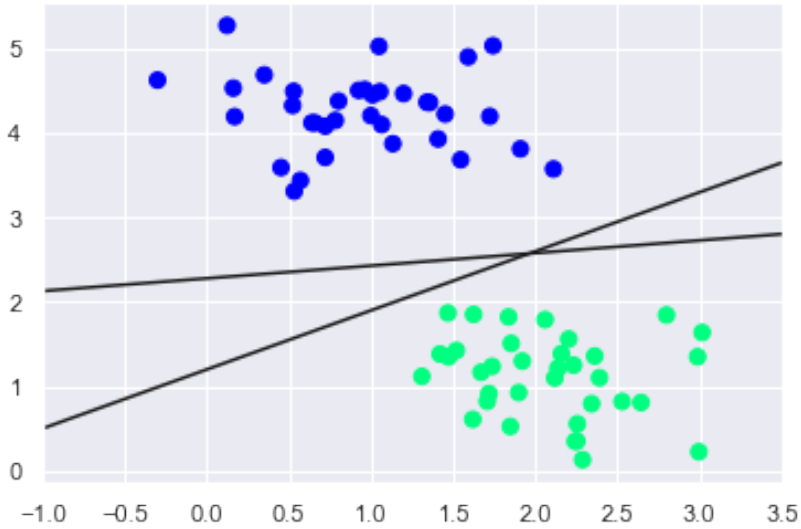
Apprendimento supervisionato

Attraverso una fase di addestramento in cui si tengono in considerazione gli output corretti, viene approssimata la funzione che mappa ogni paziente nella sua terapia.



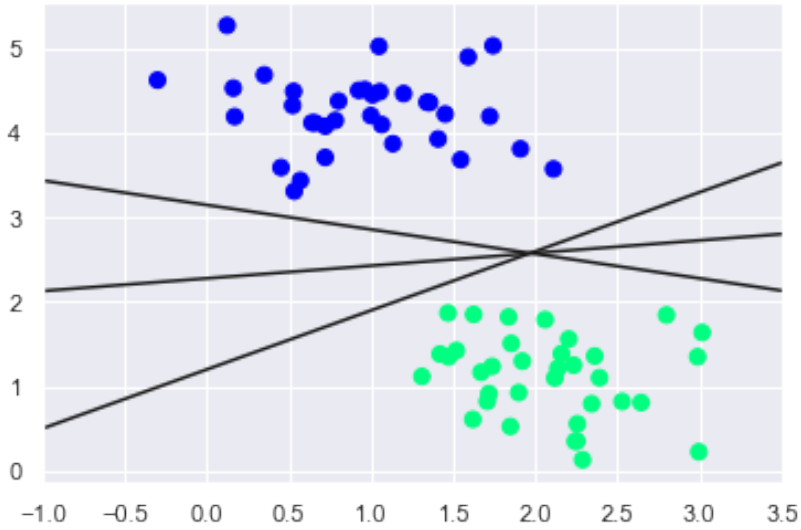
Apprendimento supervisionato

Attraverso una fase di addestramento in cui si tengono in considerazione gli output corretti, viene approssimata la funzione che mappa ogni paziente nella sua terapia.



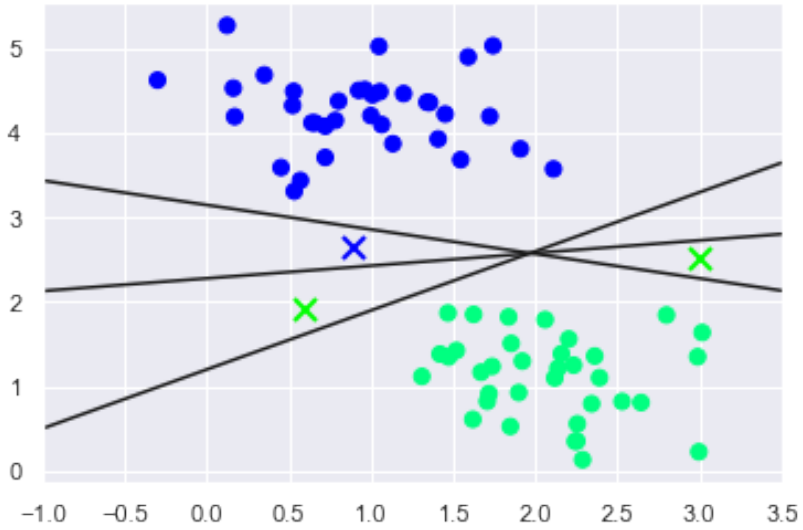
Apprendimento supervisionato

Attraverso una fase di addestramento in cui si tengono in considerazione gli output corretti, viene approssimata la funzione che mappa ogni paziente nella sua terapia.



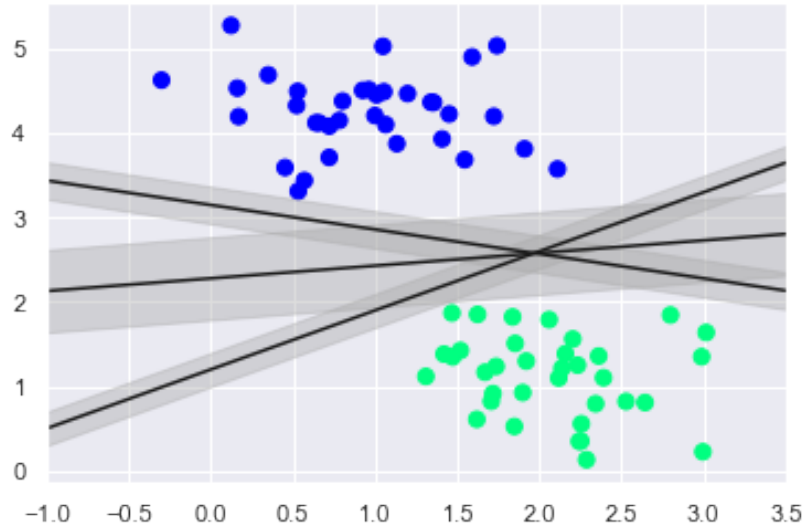
Apprendimento supervisionato

Attraverso una fase di addestramento in cui si tengono in considerazione gli output corretti, viene approssimata la funzione che mappa ogni paziente nella sua terapia.

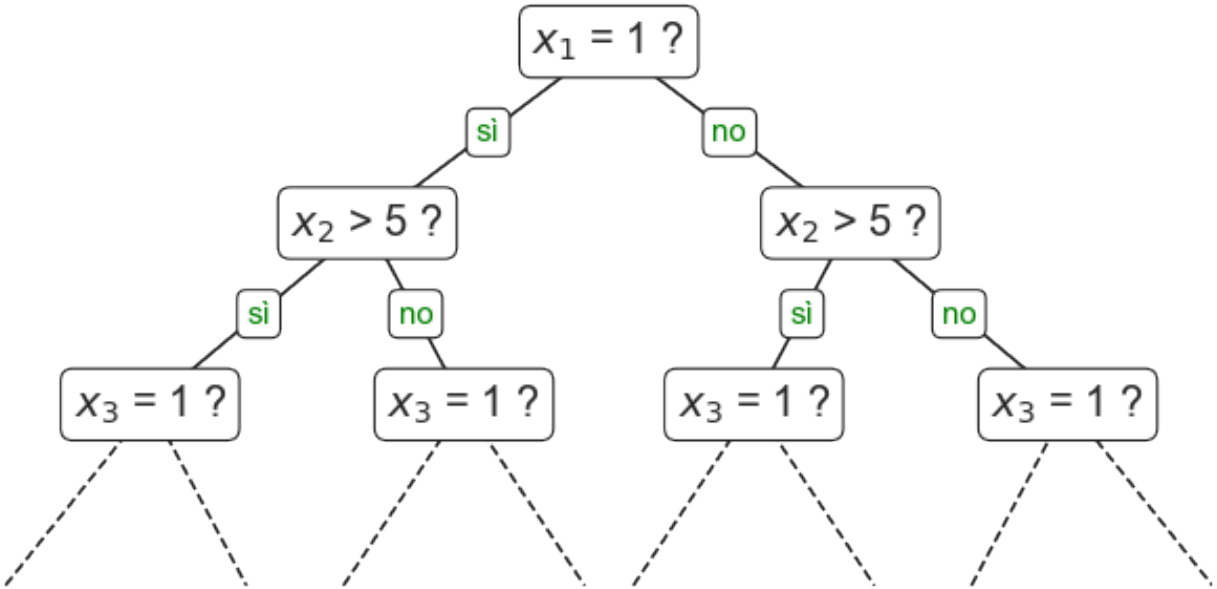


Apprendimento supervisionato

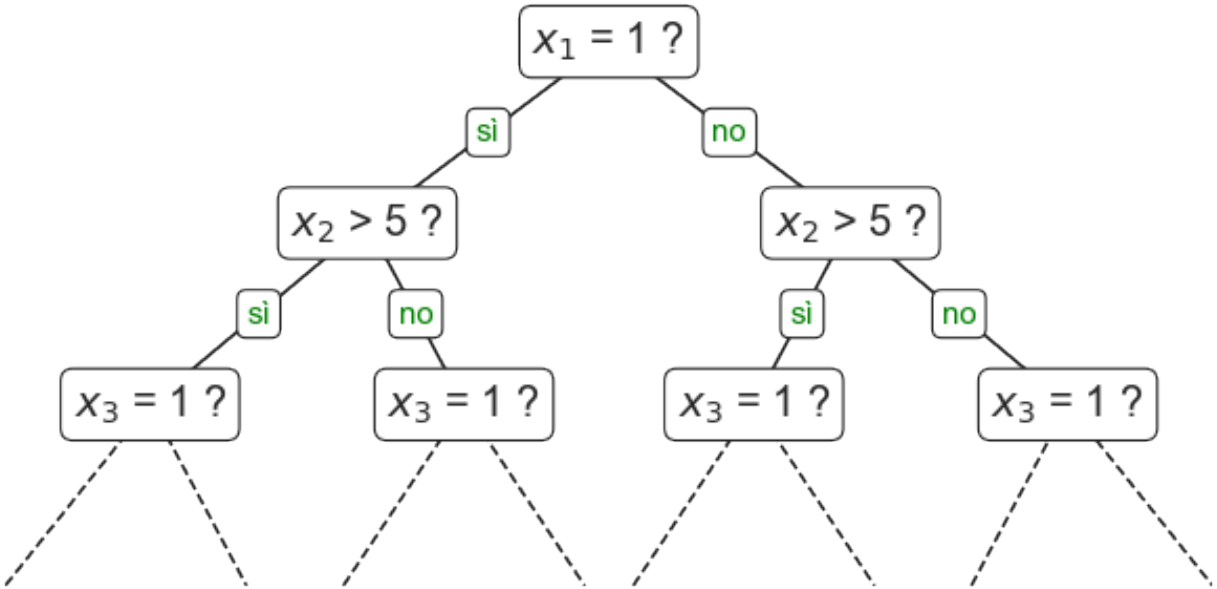
Attraverso una fase di addestramento in cui si tengono in considerazione gli output corretti, viene approssimata la funzione che mappa ogni paziente nella sua terapia.



Decision Tree: suddivisione ricorsiva del dataset

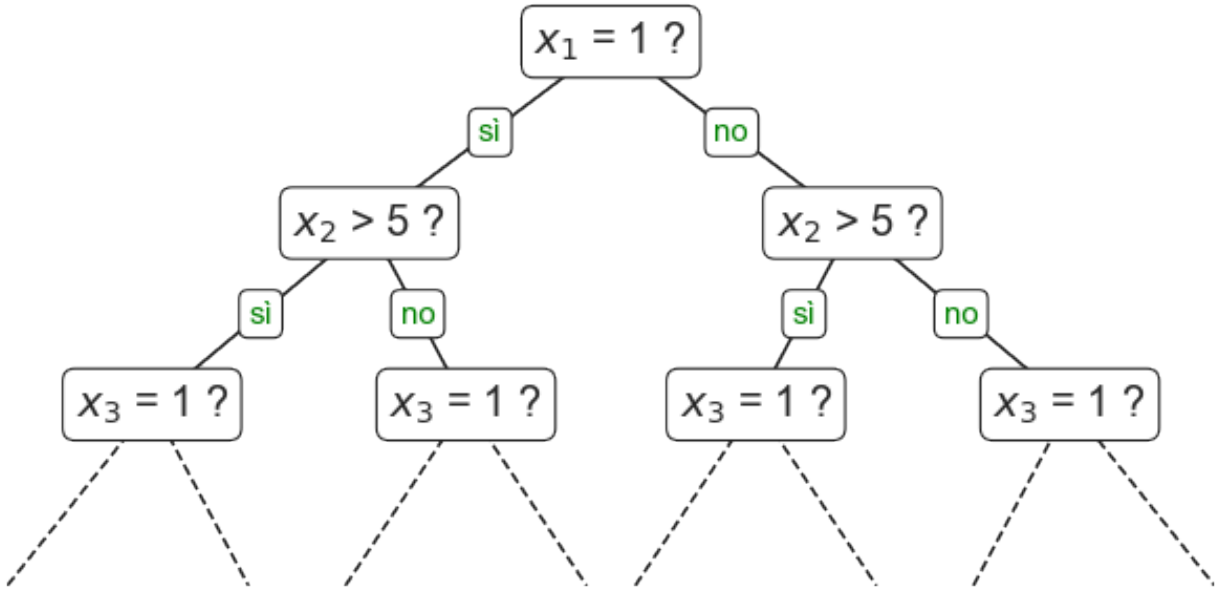


Decision Tree: suddivisione ricorsiva del dataset



Problema: overfitting

Decision Tree: suddivisione ricorsiva del dataset



Problema: overfitting

Soluzione

Random Forest: combinazione di più decision tree

Selezione Iperparametri

Valori che regolano il processo di addestramento del modello

Selezione Iperparametri

Valori che regolano il processo di addestramento del modello

- SVM: $C = 5$ e $\gamma = 0,005$

Selezione Iperparametri

Valori che regolano il processo di addestramento del modello

- SVM: $C = 5$ e $\gamma = 0,005$
- Random Forest: `n_estimators`=250 e `max_depth`=90

Selezione Iperparametri

Valori che regolano il processo di addestramento del modello

- SVM: $C = 5$ e $\gamma = 0,005$

  $\Rightarrow$ Cross validation accuracy = 86,15 %

- Random Forest: `n_estimators`=250 e `max_depth`=90

  $\Rightarrow$ Cross validation accuracy = 87,69 %

Feature selection

Strumenti coinvolti:

- Selezione delle Feature Univariata (UFS)
- Random Forest

Feature selection

Strumenti coinvolti:

- Selezione delle Feature Univariata (UFS)
- Random Forest

Feature Ranking tramite UFS

1. celebral malaria/coma (WHO)
2. jaundice (WHO)
3. hyperparasitemia (WHO)
4. parassitemia baseline %
5. anemia (WHO)

6. acute renal failure (WHO)
7. after 24 hr parassitemia baseline %
8. bil 1° giorno
9. creat 1° giorno
10. AST 1° giorno

11. shock (WHO)
12. comorbitiy
13. hypoglycaemia (WHO)
14. respiratory failure (WHO)
15. liver function test >3 time normal range (WHO)

Feature Ranking tramite Random Forest

1. bil 1° giorno
2 cerebral malaria/coma
3. parassitemia baseline %
4. PLT 1°giorno
5. Hb 1° giorno

6. ALT 1° giorno
7. età
8 after 24 hr parassitemia baseline %
9. AST 1° giorno
10. Na 1° giorno

11. durata permanenza (giorni)
12. creat 1° giorno
13. comorbitiy
14. ritardo diagnosi (gg)
15. ritardo accesso cure (gg)

Feature selection

Miglioramento dei classificatori

Risultato dei classificatori dopo la feature selection:

Classificatore	**Feature Ranking**	**Feature considerate**	**Valore degli Iperparametri**	**Accuracy**
SVM	UFS	7	C = 1 gamma = 0,25	92,31 %
SVM	Random Forest	7	C = 2 gamma = 0,0625	89,23 %
Random Forest	UFS	17	n_estimators = 200 max_depth = 50	89,23 %
Random Forest	Random Forest	5	n_estimators = 200 max_depth = 50	90,77 %

Apprendimento non supervisionato

Si cerca di individuare una struttura intrinseca al dataset, senza considerare il target vector

Apprendimento non supervisionato

Si cerca di individuare una struttura intrinseca al dataset, senza considerare il target vector

Clustering: partizione del dataset in gruppi (cluster) in base alle analogie tra i campioni

Apprendimento non supervisionato

Si cerca di individuare una struttura intrinseca al dataset, senza considerare il target vector

Clustering: partizione del dataset in gruppi (cluster) in base alle analogie tra i campioni

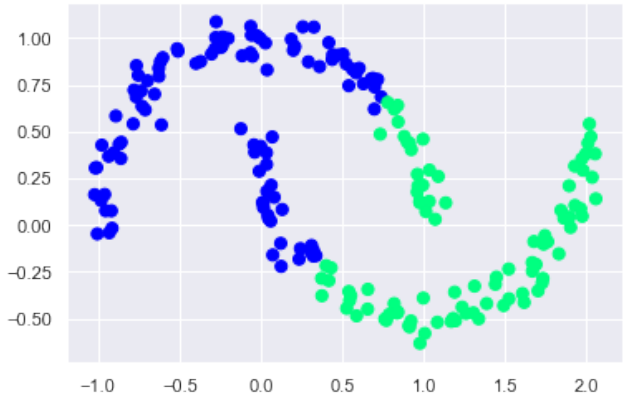
- **K-Means**

Apprendimento non supervisionato

Si cerca di individuare una struttura intrinseca al dataset, senza considerare il target vector

Clustering: partizione del dataset in gruppi (cluster) in base alle analogie tra i campioni

- **K-Means**

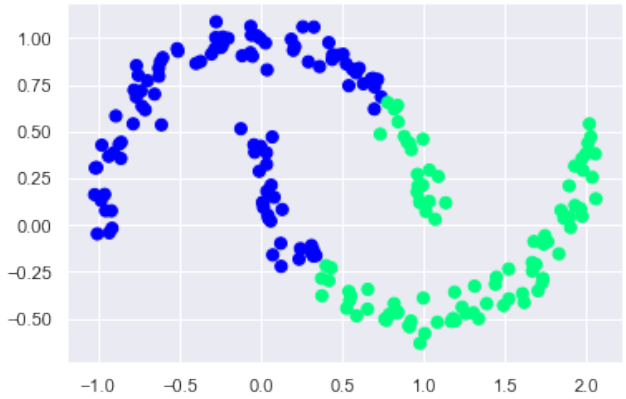


Apprendimento non supervisionato

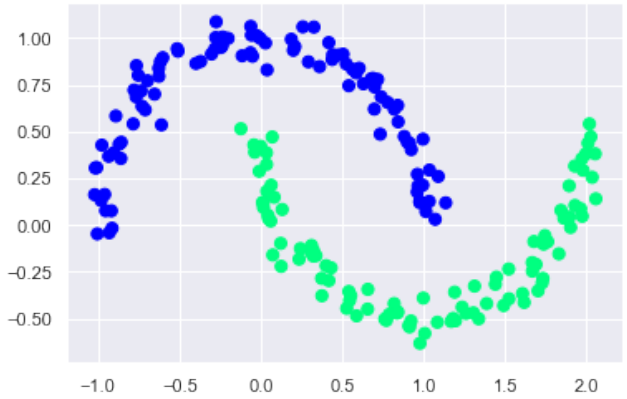
Si cerca di individuare una struttura intrinseca al dataset, senza considerare il target vector

Clustering: partizione del dataset in gruppi (cluster) in base alle analogie tra i campioni

- **K-Means**



- **Spectral Clustering**



Apprendimento non supervisionato

Risultati clustering

Algoritmo di Clustering	K-Means	Spectral Clustering	K-Means	Spectral Clustering
Strumento di Feature Selection	PCA	PCA	sPCA	sPCA
Numero di pazienti nel cluster più grande	95	67	114	74
di cui con terapia orale	60	44	65	46
Numero di pazienti nel cluster più piccolo	24	52	5	45
di cui con terapia orale	9	25	4	23

Tabella: Risultati Clustering preceduto da una feature selection.

Conclusione

Risultati e commenti finali

- Costruzione di classificatori con buoni livelli di accuratezza con cui scegliere la terapia nei casi futuri

Conclusione

Risultati e commenti finali

- Costruzione di classificatori con buoni livelli di accuratezza con cui scegliere la terapia nei casi futuri
- Individuazione delle feature più rilevanti ai fini della selezione della terapia

Conclusione

Risultati e commenti finali

- Costruzione di classificatori con buoni livelli di accuratezza con cui scegliere la terapia nei casi futuri
- Individuazione delle feature più rilevanti ai fini della selezione della terapia
- Studio delle sovrapposizioni tra i cluster ottenuti per un'analisi più dettagliata da parte dei medici

Conclusione

Vi ringrazio per l'attenzione