Data Analytics project

Daily and Sports Activities Data Set

Gruppo 14
- Costanzo Giuseppe
- Masci Mattia
- Tagliente Gabriele

Sommario

- Introduzione al dataset
- Data cleaning e descrizione del dataset
- Exploratory analysis
- Unsupervised learning (clustering)
- Supervised learning (decision tree e svm)
- Risultati

Introduzione del dataset

- 19 attività svolte da 8 soggetti (4 maschi e 4 femmine)
  - sitting (A1),
    standing (A2),
    lying on back and on right side (A3 and A4),
    ascending and descending stairs (A5 and A6),
    standing in an elevator still (A7)
    and moving around in an elevator (A8),
    walking in a parking lot (A9),
    walking on a treadmill with a speed of 4 km/h (in flat and 15 deg inclined positions) (A10 and A11),
    running on a treadmill with a speed of 8 km/h (A12),
    exercising on a stepper (A13),
    exercising on a cross trainer (A14),
    cycling on an exercise bike in horizontal and vertical positions (A15 and A16),
    rowing (A17),
    jumping (A18),
    and playing basketball (A19)
- L'attività dura 5 minuti
- 5 unità sul corpo
- 3 tipologie di sensori (x unità) : Accelerometro, Giroscopio e Magnetometro (tri-assiali)

Data cleaning e descrizione del dataset

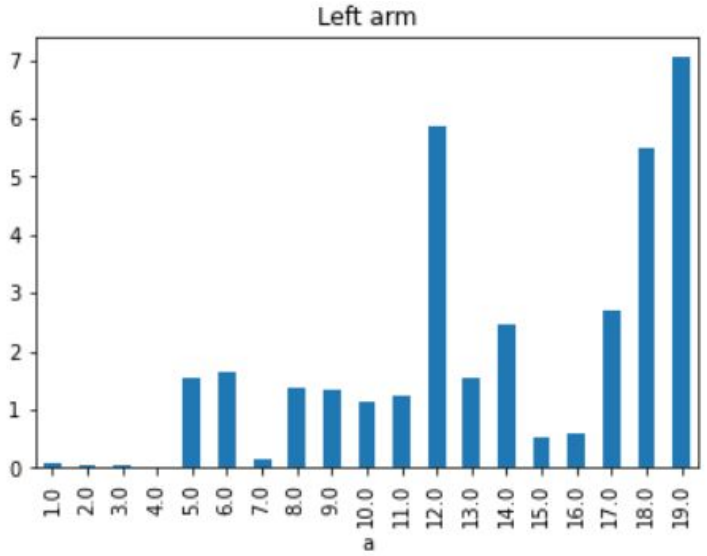
- 19 attività x 8 soggetti x 60 segmenti
- Campionamento a 25 Hz
- Ogni segmento
  - 125 righe : 5 sec x 25 Hz
  - 45 colonne : 5 unità x 9 assi totali
- Ogni colonna contiene 125 campioni di dati da uno dei sensori delle unità per un periodo di 5 secondi
- Ogni riga contiene i dati acquisiti da tutti i sensori (45 valori) in un particolare istante di campionamento
- Il dataset in origine non contiene valori nulli
- Il dataset finale manipolato contiene :
  - 9120 righe (19 attività x 8 soggetti x 60 segmenti) (da 1 140 000)
  - 315 colonne (45 assi x 7 valori considerati) (da 45)

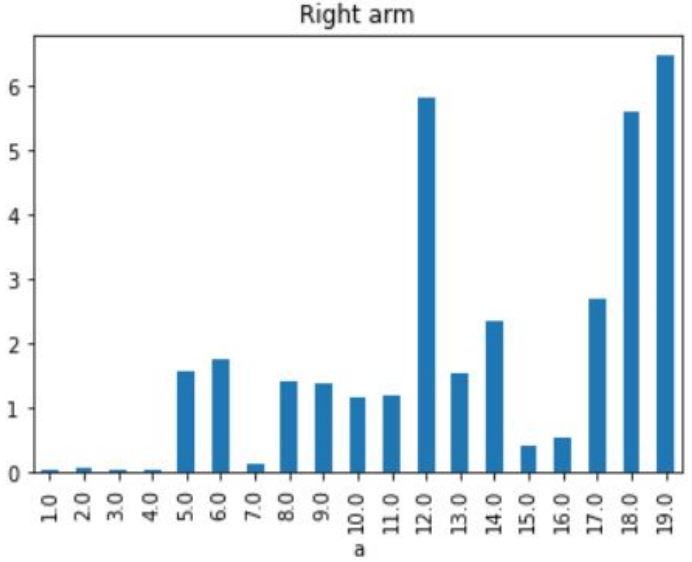
Data cleaning e descrizione del dataset

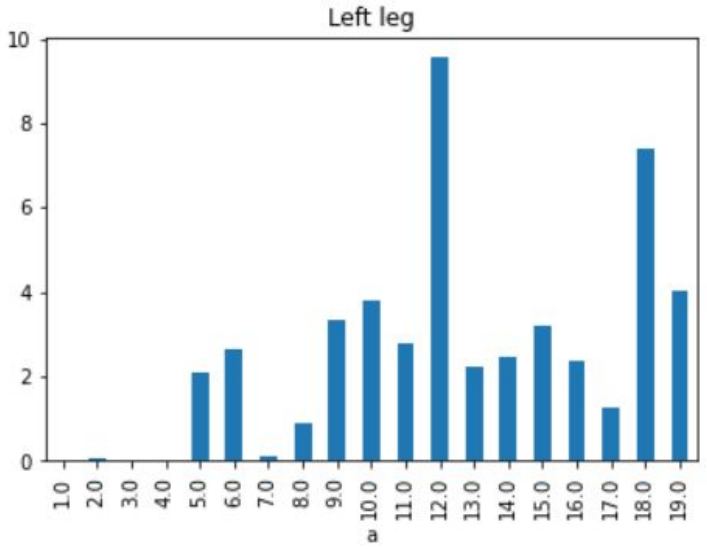
- Vengono considerate (per ogni singolo asse dei sensori)
  - Mean
  - Variance
  - Min
  - Max
  - Quartili (25, 50, 75)

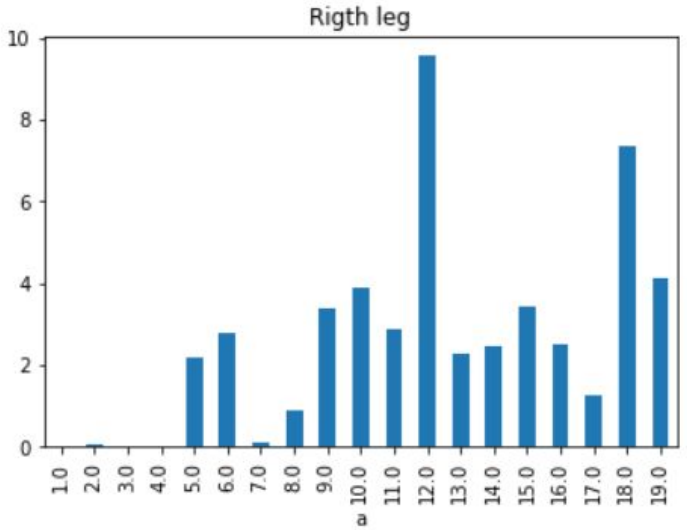
	(T, xacc, mean)	(T, xacc, std)	(T, xacc, min)	(T, xacc, 25%)	(T, xacc, 50%)	(T, xacc, 75%)	(T, xacc, max)	(T, yacc, mean)	(T, yacc, std)	(T, yacc, min)	...	(LL, ymag, 50%)	(LL, ymag, 75%)	(LL, ymag, max)	(LL, zmag, mean)	(LL, zmag, std)
0	9.375416	1.340390	7.2316	8.40890	9.0251	10.0480	12.995	-2.220890	0.841599	-3.7353	...	0.41325	0.51501	0.59019	-0.321324	0.039035
1	9.305754	1.284732	7.4804	8.41810	9.0577	9.7066	13.283	-2.195375	0.841058	-3.8709	...	0.28580	0.44353	0.55931	-0.414925	0.041983
2	9.304878	1.106255	7.7173	8.41870	9.1712	9.7528	12.333	-2.062480	0.886773	-3.9881	...	0.28769	0.43510	0.53405	-0.390564	0.038960
3	9.362854	1.362537	7.2450	8.42340	8.9843	9.9633	13.152	-2.198233	0.794485	-3.8168	...	0.38984	0.45443	0.54418	-0.261211	0.038384
4	9.174813	1.237477	7.0816	8.28350	8.9143	9.8522	12.738	-2.587231	0.875247	-4.5112	...	0.32838	0.42768	0.52017	-0.208207	0.030420
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

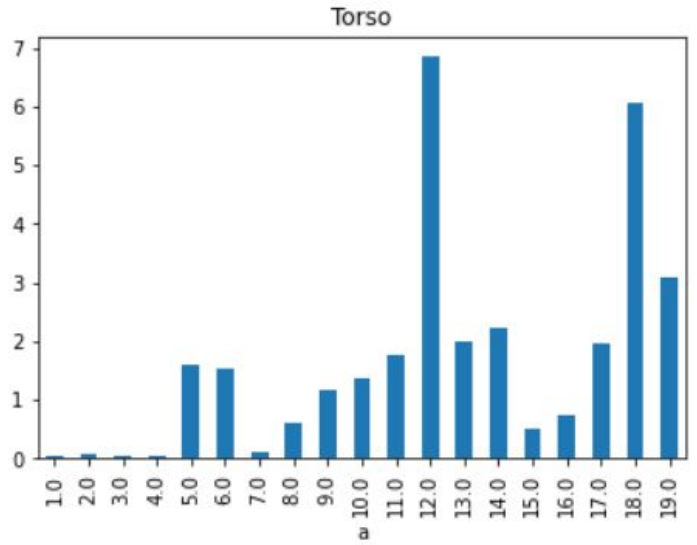
Data cleaning e descrizione del dataset - Accelerometro





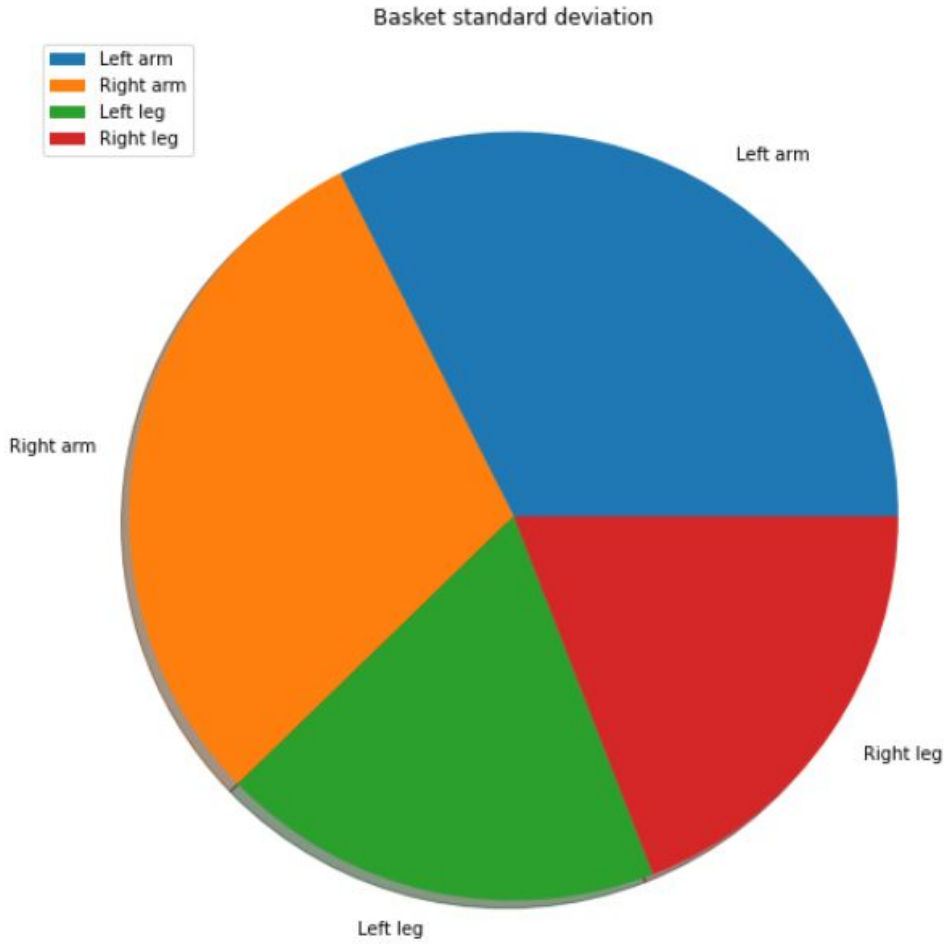


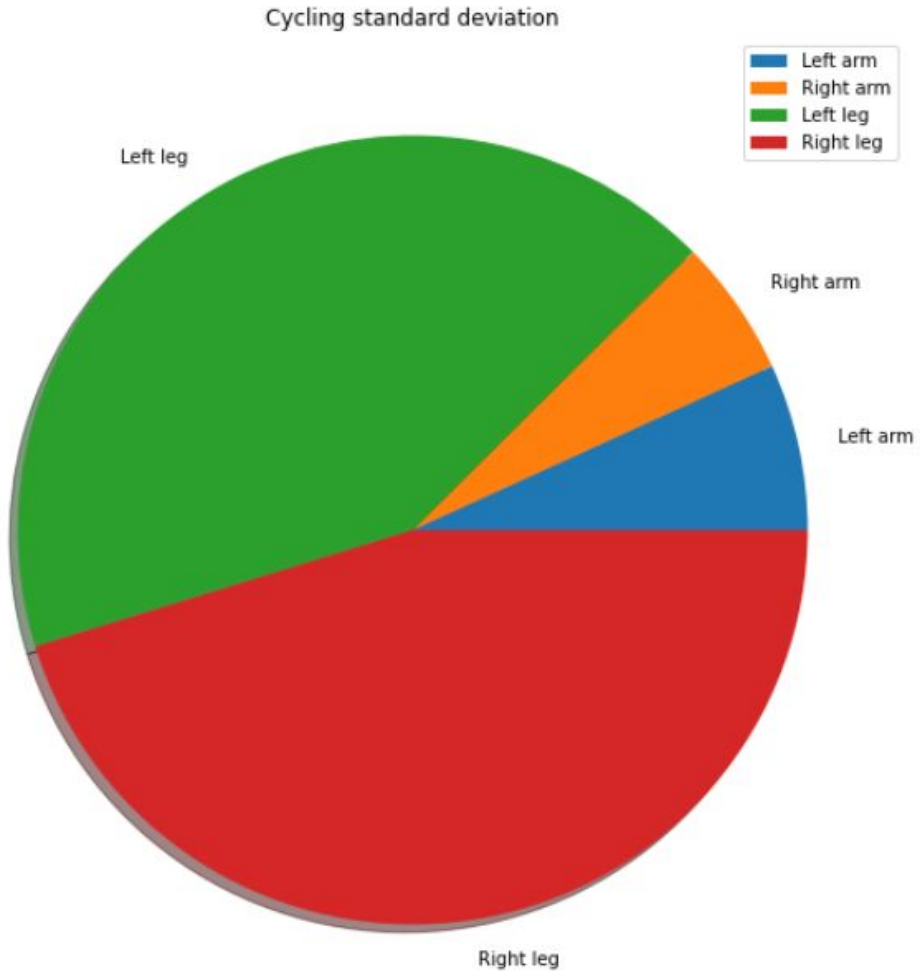




Data cleaning e descrizione del dataset

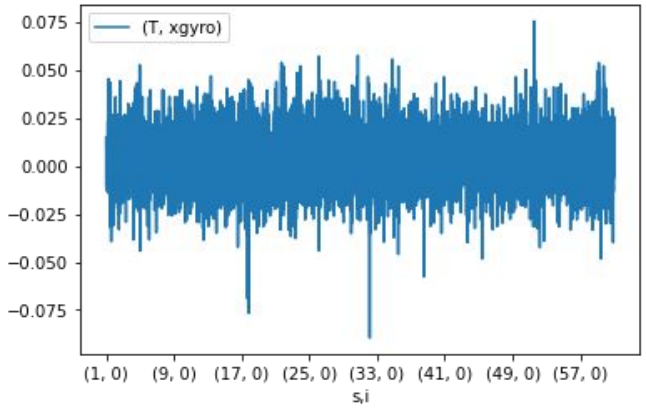
- Deviazione standard riferita a varie parti del corpo in diverse attività



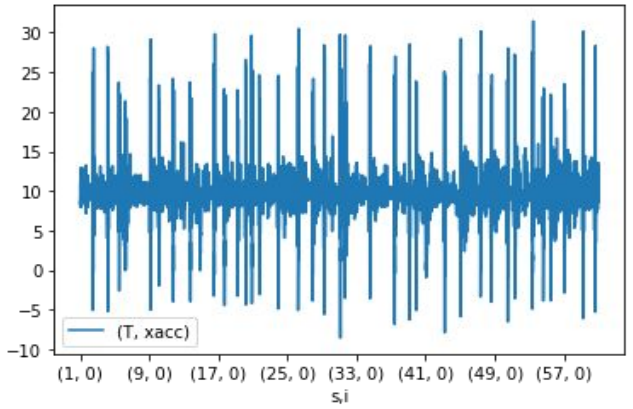


Data cleaning e descrizione del dataset

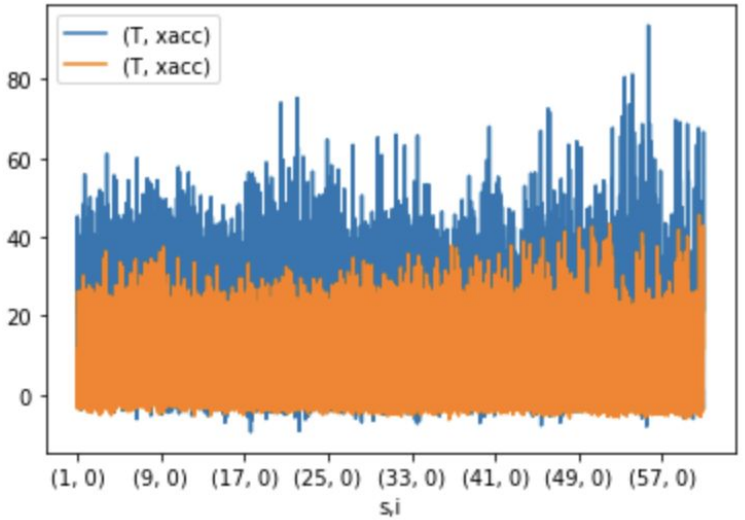
Sit



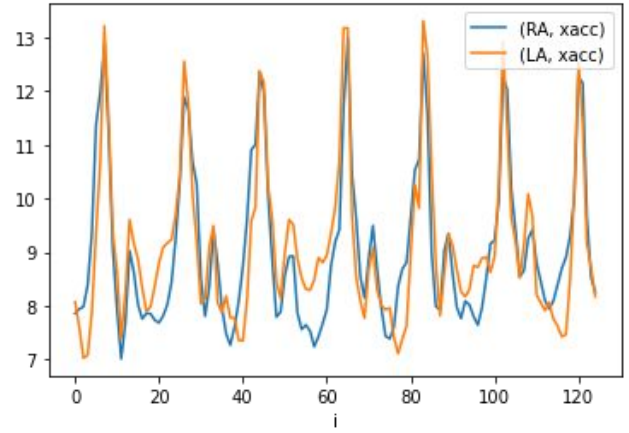
Basket



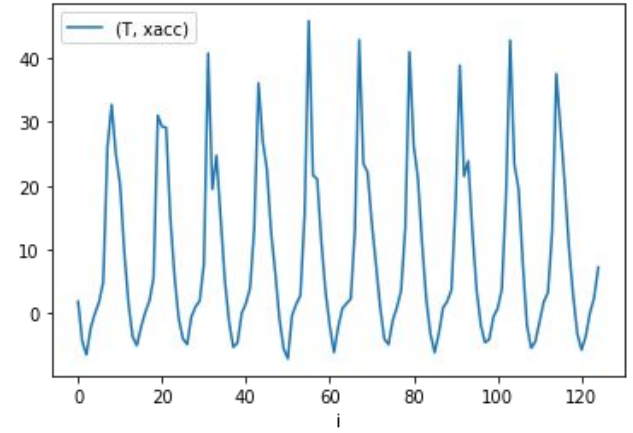
Running



Walk

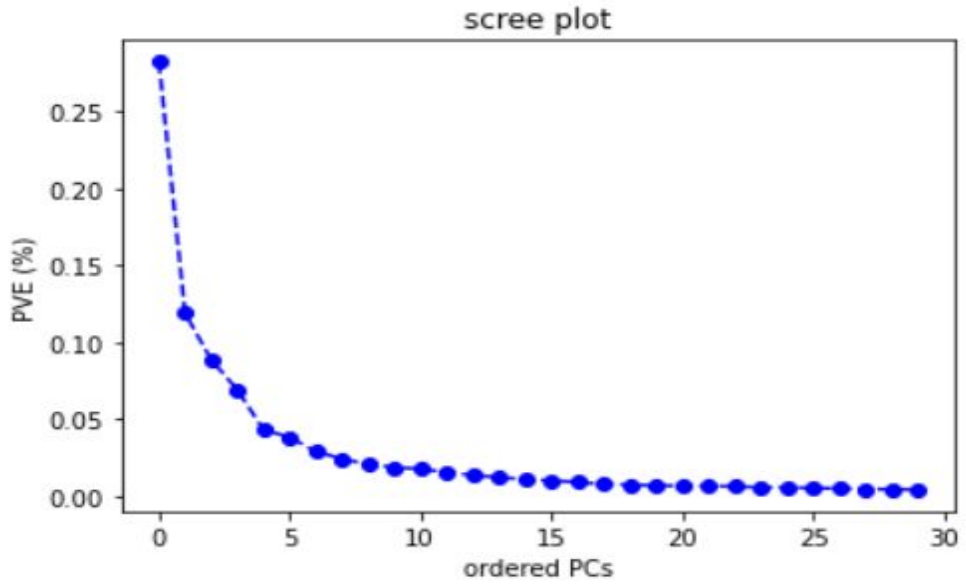


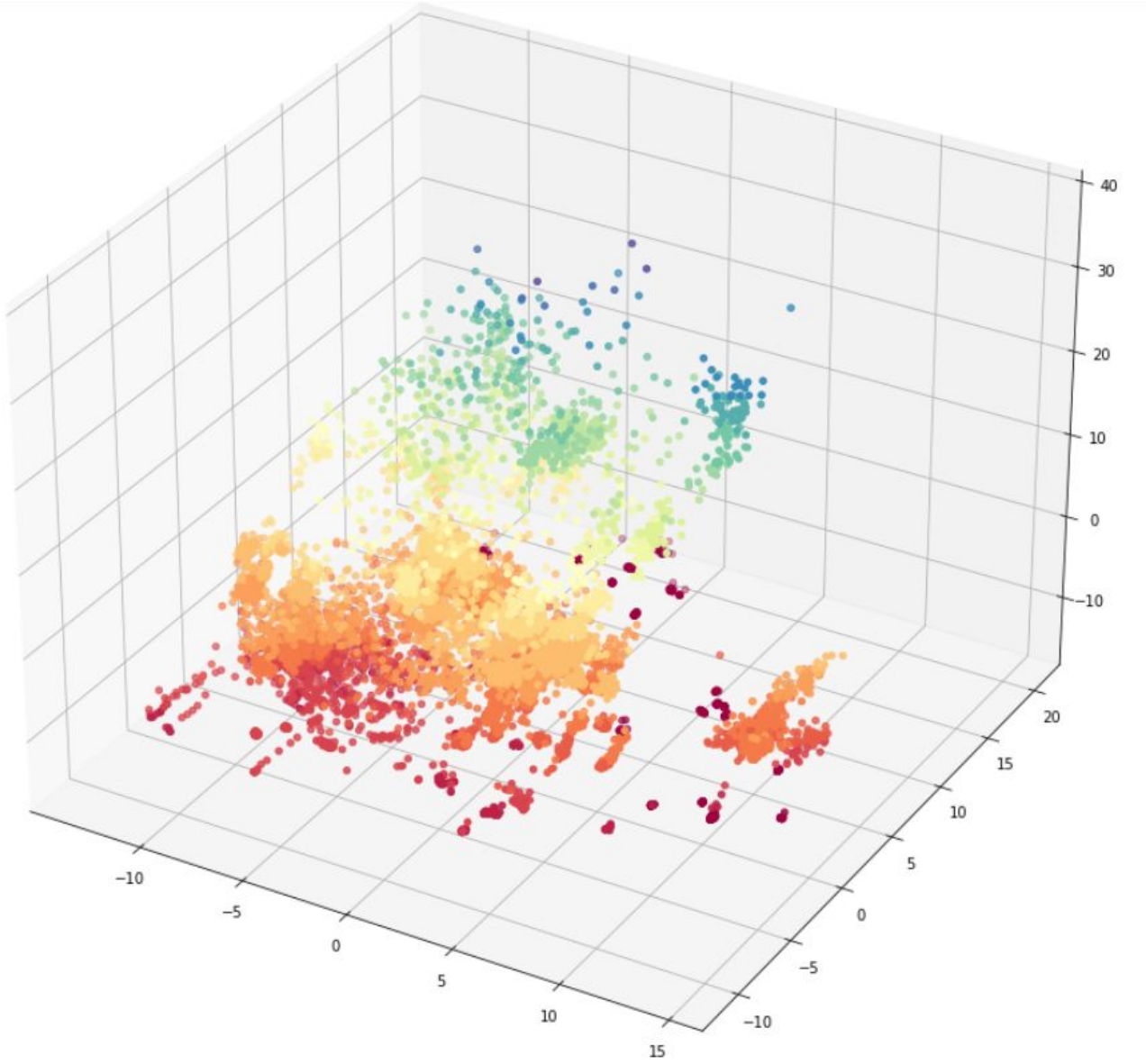
Jump



Exploratory Analysis

- PCA
  - Scaling dei dati
  - Fit

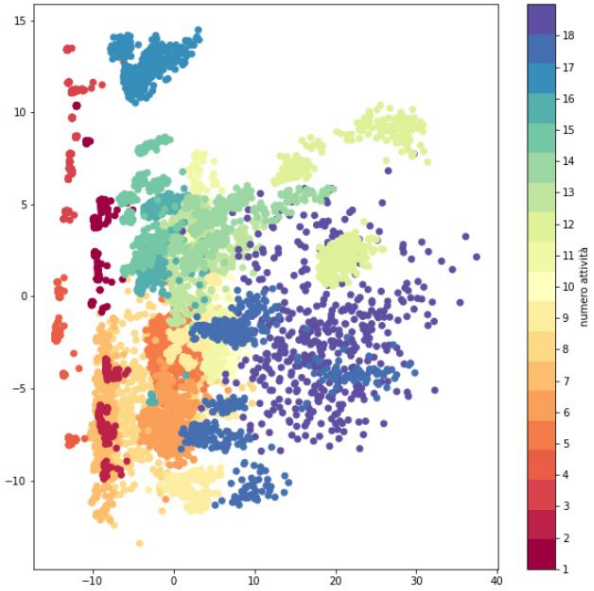


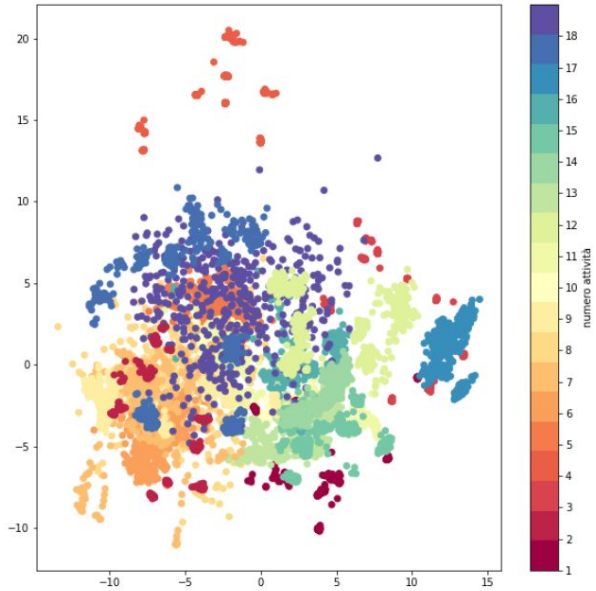


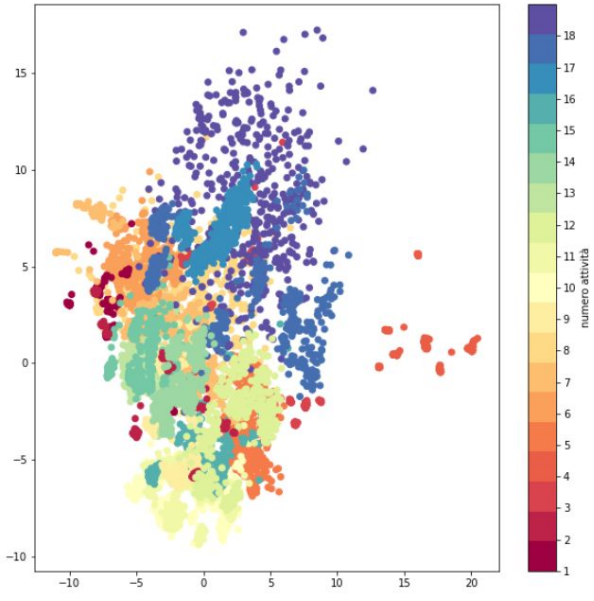
PCA

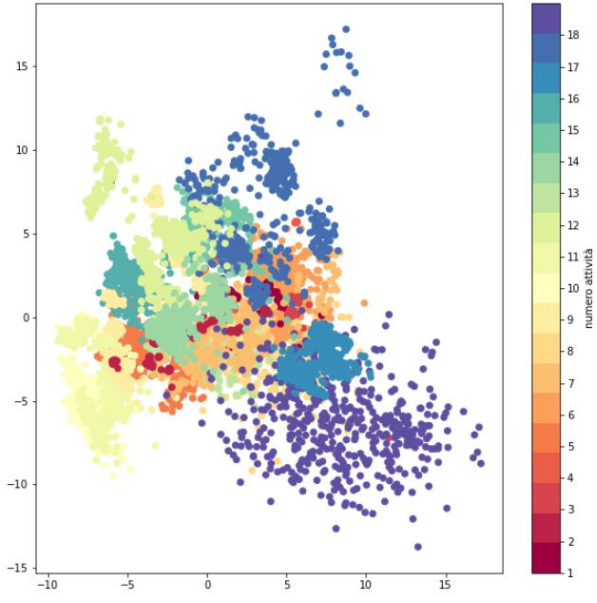
(1,2) | (2,3)

(3,4) | (4,5)



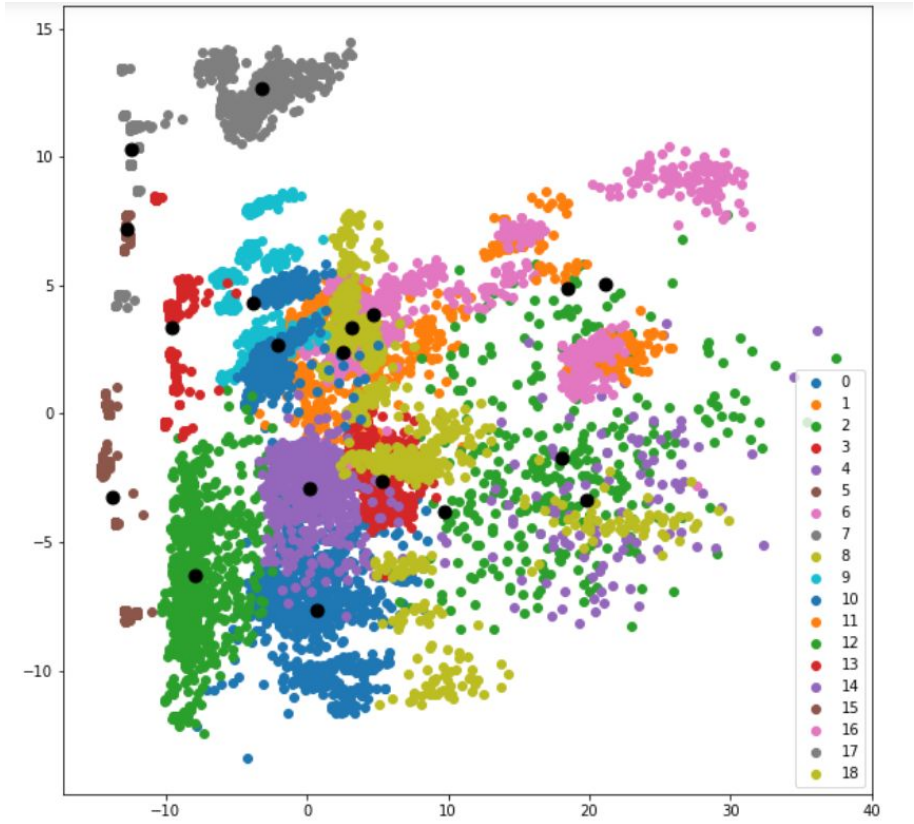




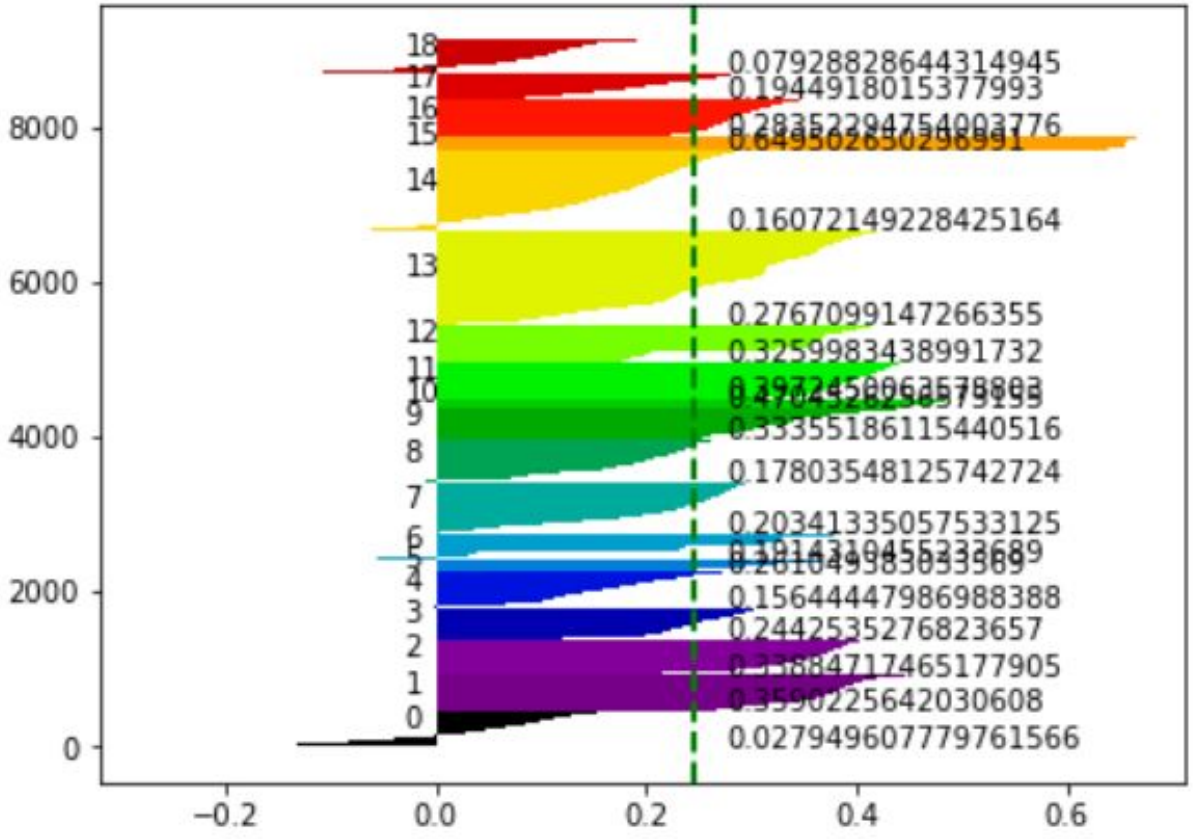


Unsupervised Learning - clustering e silhouette

- K-Means K=19



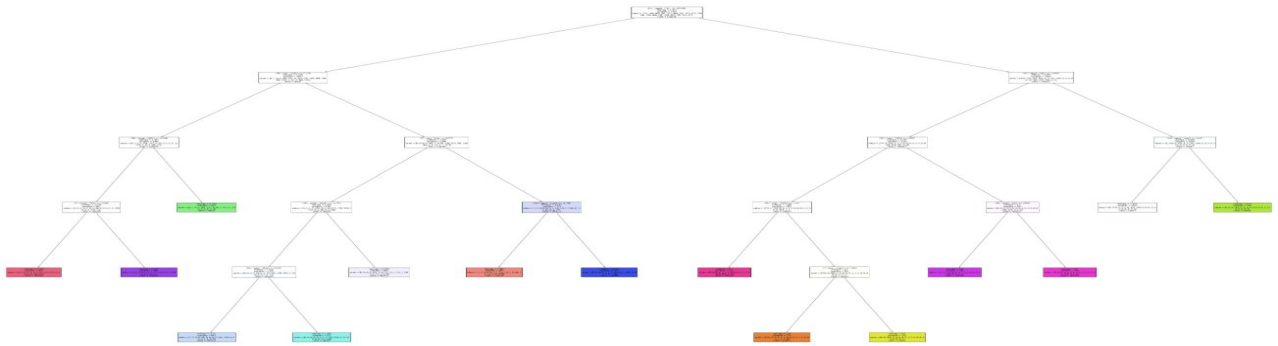

```
overall clustering silhouette 0.2463097416525351

<matplotlib.collections.PathCollection at 0x2528838cfd0>
```

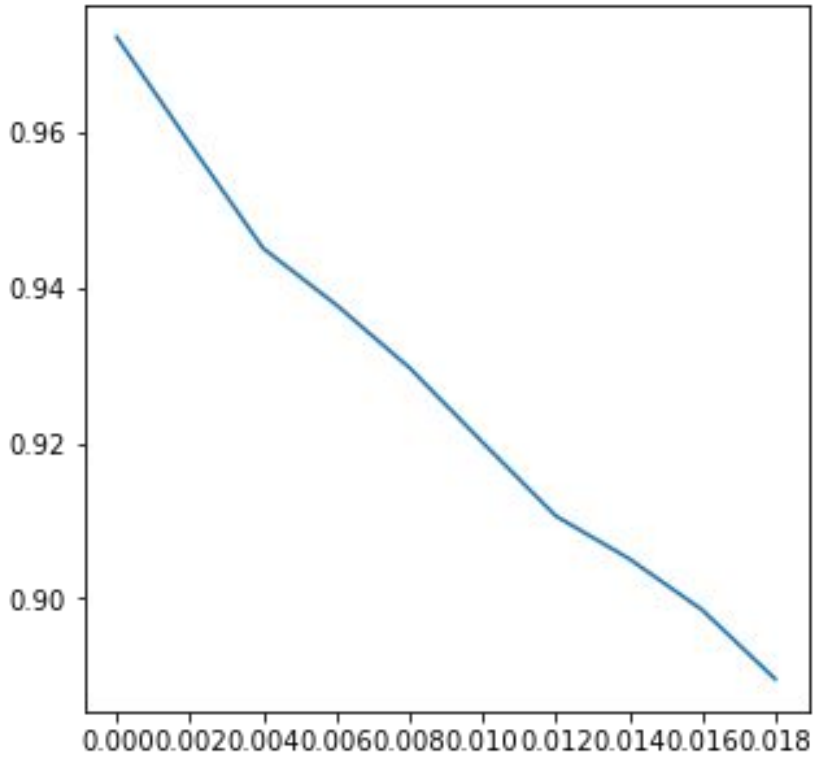


Supervised Learning - Decision Tree

- effettuato dei test senza e con pruning
- calcolato il miglior alfa tramite grid-search utilizzando 5-fold per la cross-validation.

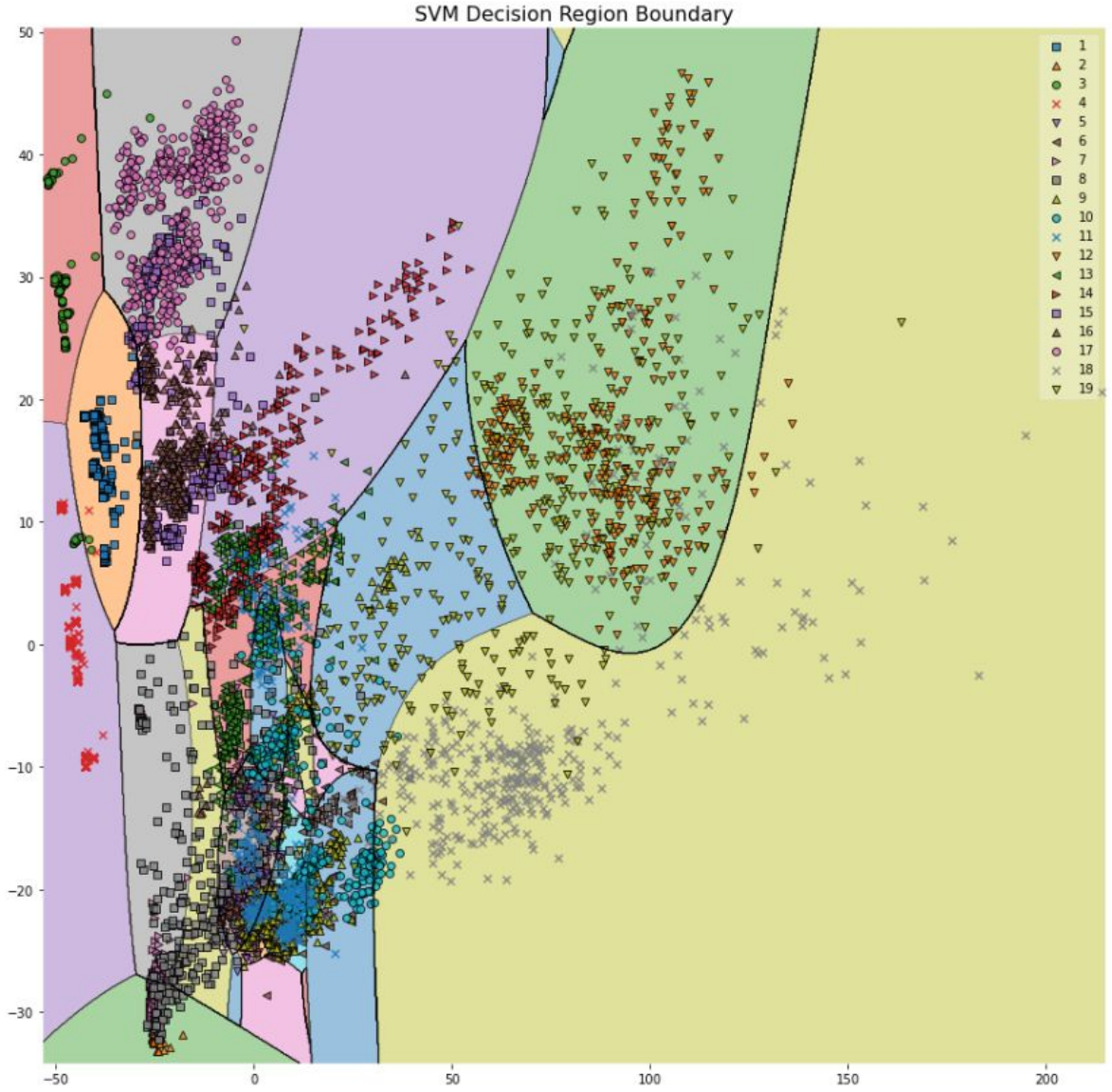
alpha	# levels	# leaves	testing accuracy
0.0	15	111	0.985
0.1	5	15	0.745



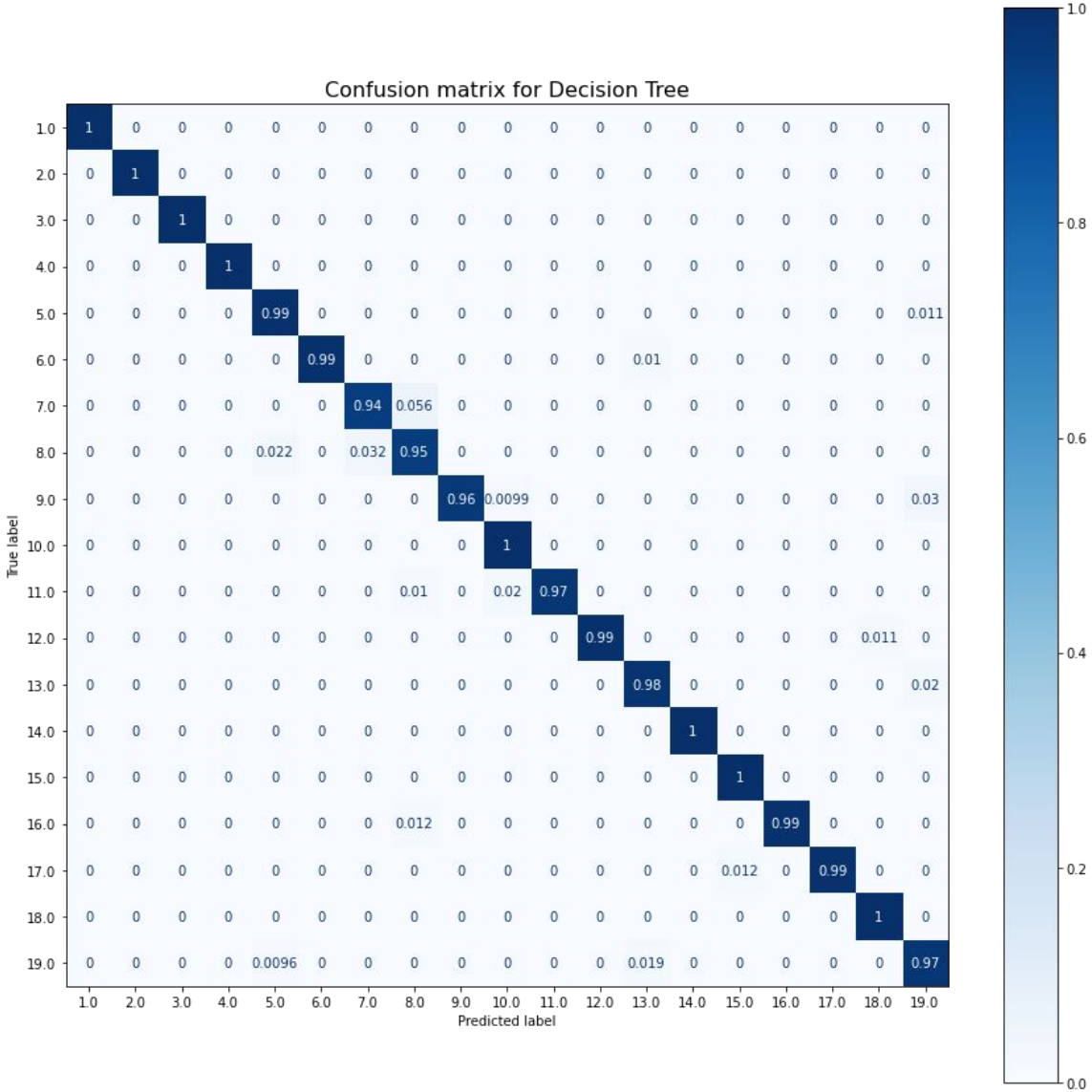


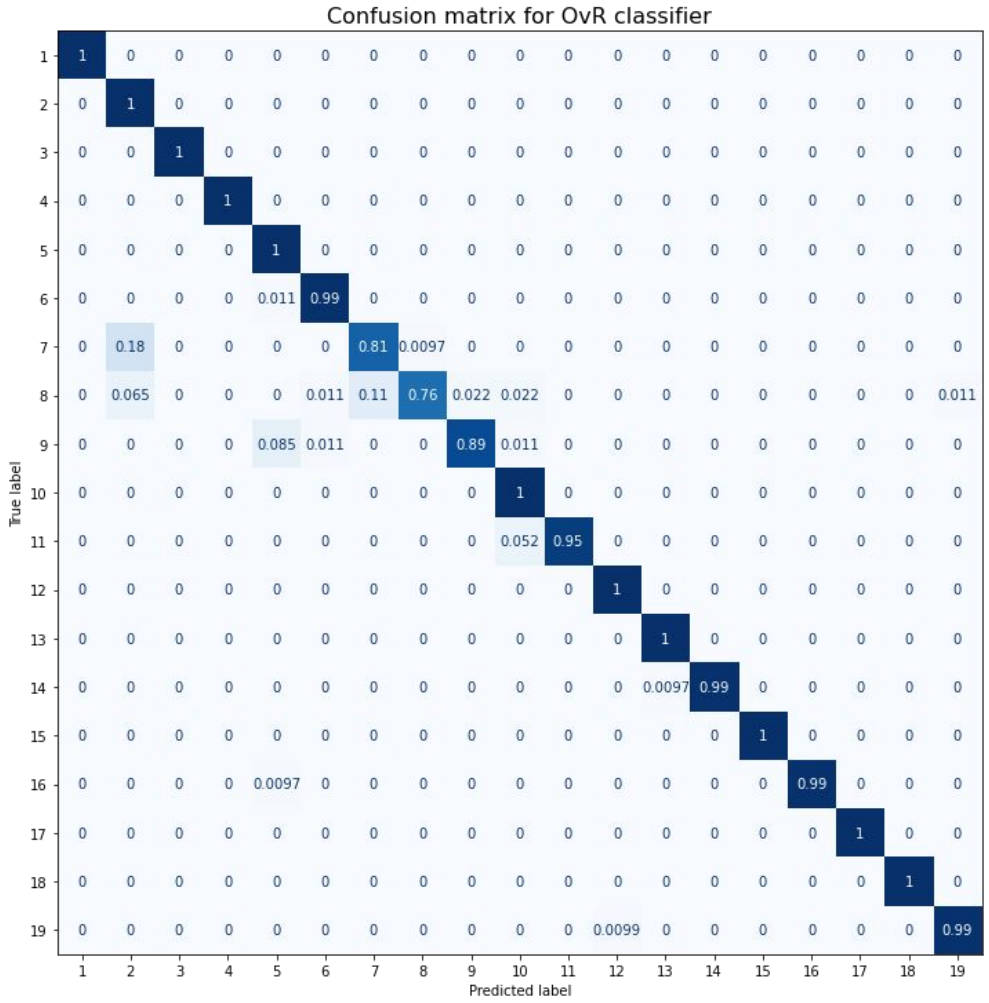
SVM - OvR

- Support Vector Machine (nativamente binario)
- reso multiclasse con tecnica One vs Rest
  - un classificatore binario per ogni classe
  - restituisce una probabilità di appartenere ad una classe
  - risultato: probabilità più alta.



Risultati - confusion matrices





Risultati

- 2 modelli di Supervised Learning: Decision Tree e SVM
- Decision Tree ha un'accuratezza migliore di SVM, anche se pure SVM dà buoni risultati

	Training	Testing	Training BA	Testing BA
Decision Tree	1.000000	0.985197	1.000000	0.985024
SVM	0.972725	0.966557	0.972684	0.966713