

MYOCARDIAL INFARCTION- COMPLICATIONS ANALYSIS

DATA MINING : CASO DI STUDIO
VITO SIMONE LACATENA

MYOCARDIAL INFARCTION-COMPLICATIONS

BUSINESS UNDERSTANDING

BACKGROUND

L'IM è uno dei problemi più impegnativi della medicina moderna. L'infarto miocardico acuto è associato ad un'elevata mortalità nel primo anno successivo. L'incidenza di IM rimane elevata in tutti i paesi. Ciò è particolarmente vero per la popolazione urbana dei paesi altamente sviluppati, esposta a fattori di stress cronico, alimentazione irregolare e non sempre equilibrata. Negli Stati Uniti, ad esempio, ogni anno più di un milione di persone soffrono di infarto del miocardio e 200-300mila muoiono di infarto miocardico acuto prima di arrivare in ospedale.

Il decorso della malattia nei pazienti con infarto miocardico è diverso. L'infarto del miocardio può verificarsi senza complicazioni o con complicanze che non peggiorano la prognosi a lungo termine. Allo stesso tempo, circa la metà dei pazienti nei periodi acuto e subacuto presenta complicazioni che portano al peggioramento della malattia e persino alla morte. Anche uno specialista esperto non può sempre prevedere lo sviluppo di queste complicazioni. A questo proposito, la previsione delle complicanze dell'infarto miocardico al fine di attuare tempestivamente le necessarie misure preventive è un compito importante.

OBBIETTIVI DI BUSINESS

Fenotipizzazione della malattia: Individuazione di sotto-gruppi di pazienti sulla base delle informazioni possedute, e determinare l'esistenza di una possibile relazione tra le complicazioni della malattia che i pazienti presentano nel corso della malattia.

OBBIETTIVI DI DATA MINING

1. Analisi dei Cluster, individuare possibili sottogruppi di pazienti mediante algoritmi di clustering.
2. Association Rules Extraction, estrazione di regole di associazione per estrarre relazioni nascoste dai dati

DATA UNDERSTANDING

DATASET

Il dataset utilizzato è il **Myocardial infarction complications Data Set** reperibile nella repository <https://archive.ics.uci.edu/ml/datasets/Myocardial+infarction+complications>

INFORMAZIONI GENERALI

Ospedale in cui sono stati raccolti i dati: Krasnoyarsk Interdistrict Clinical Hospital №20 intitolato a I. S. Berzon (Russia)

Periodo di raccolta dati: 1992-1995

N.ro osservazioni: 1700

N.ro Feature: 123 di cui 111 variabili di input e 12 possibili complicazioni utilizzabili come variabili di output

Ci sono quattro possibili momenti temporali per la previsione della complicazione: sulla base delle informazioni note:

Feature di input misurata prima della fine del primo giorno (24 ore dopo il ricovero in ospedale)

R_AB_1_n, NA_R_1_n, NOT_NA_1_n

Feature di input misurata prima della fine del secondo giorno (48 ore dopo il ricovero in ospedale)

R_AB_2_n, NA_R_2_n, NOT_NA_2_n

Feature di input misurata prima della fine del terzo giorno (72 ore dopo il ricovero in ospedale)

R_AB_3_n, NA_R_3_n, NOT_NA_3_n

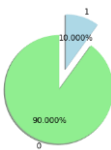
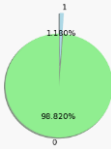
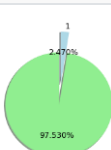
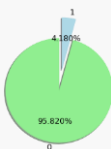
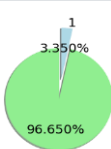
Feature di input misurata al momento dell'ammissione in ospedale

Tutte le altre feature di input.

VARIABILI DI OUTPUT (TIPI DI COMPLICAZIONI)

BINARIE

MYOCARDIAL INFARCTION-COMPLICATIONS

Nome	Descrizione	Diagramma
FIBR_PREDS	Atrial fibrillation	 <p>A pie chart representing the distribution of Atrial fibrillation. The chart is divided into two segments: a large green segment representing 90.000% and a smaller blue segment representing 10.000%. The numbers 0 and 1 are visible near the segments.</p>
PREDS_TAH	Supraventricular tachycardia	 <p>A pie chart representing the distribution of Supraventricular tachycardia. The chart is divided into two segments: a large green segment representing 96.820% and a smaller blue segment representing 3.180%. The numbers 0 and 1 are visible near the segments.</p>
JELUD_TAH	Ventricular tachycardia	 <p>A pie chart representing the distribution of Ventricular tachycardia. The chart is divided into two segments: a large green segment representing 97.530% and a smaller blue segment representing 2.470%. The numbers 0 and 1 are visible near the segments.</p>
FIBR_JELUD	Ventricular fibrillation	 <p>A pie chart representing the distribution of Ventricular fibrillation. The chart is divided into two segments: a large green segment representing 95.820% and a smaller blue segment representing 4.180%. The numbers 0 and 1 are visible near the segments.</p>
A_V_BLOK)	Third-degree AV block	 <p>A pie chart representing the distribution of Third-degree AV block. The chart is divided into two segments: a large green segment representing 96.650% and a smaller blue segment representing 3.350%. The numbers 0 and 1 are visible near the segments.</p>

Nome	Descrizione	Diagramma									
OTEK_LANC	Pulmonary edema	 <table border="1"><thead><tr><th>Category</th><th>Value</th><th>Percentage</th></tr></thead><tbody><tr><td>Green</td><td>90</td><td>90.650%</td></tr><tr><td>Blue</td><td>9</td><td>9.350%</td></tr></tbody></table>	Category	Value	Percentage	Green	90	90.650%	Blue	9	9.350%
Category	Value	Percentage									
Green	90	90.650%									
Blue	9	9.350%									
RAZRIV	Myocardial rupture	 <table border="1"><thead><tr><th>Category</th><th>Value</th><th>Percentage</th></tr></thead><tbody><tr><td>Green</td><td>96</td><td>96.820%</td></tr><tr><td>Blue</td><td>3</td><td>3.180%</td></tr></tbody></table>	Category	Value	Percentage	Green	96	96.820%	Blue	3	3.180%
Category	Value	Percentage									
Green	96	96.820%									
Blue	3	3.180%									
DRESSLER	Dressler syndrome	 <table border="1"><thead><tr><th>Category</th><th>Value</th><th>Percentage</th></tr></thead><tbody><tr><td>Green</td><td>95</td><td>95.590%</td></tr><tr><td>Blue</td><td>4</td><td>4.410%</td></tr></tbody></table>	Category	Value	Percentage	Green	95	95.590%	Blue	4	4.410%
Category	Value	Percentage									
Green	95	95.590%									
Blue	4	4.410%									
ZSN	Chronic heart failure	 <table border="1"><thead><tr><th>Category</th><th>Value</th><th>Percentage</th></tr></thead><tbody><tr><td>Green</td><td>76</td><td>76.820%</td></tr><tr><td>Blue</td><td>23</td><td>23.180%</td></tr></tbody></table>	Category	Value	Percentage	Green	76	76.820%	Blue	23	23.180%
Category	Value	Percentage									
Green	76	76.820%									
Blue	23	23.180%									
REC_IM	Relapse of the myocardial infarction	 <table border="1"><thead><tr><th>Category</th><th>Value</th><th>Percentage</th></tr></thead><tbody><tr><td>Green</td><td>90</td><td>90.650%</td></tr><tr><td>Blue</td><td>9</td><td>9.350%</td></tr></tbody></table>	Category	Value	Percentage	Green	90	90.650%	Blue	9	9.350%
Category	Value	Percentage									
Green	90	90.650%									
Blue	9	9.350%									

MYOCARDIAL INFARCTION-COMPLICATIONS

VARIABILI QUALITATIVE (CATEGORICHE NOMINALI)

Nome	Descrizione	Valori Assunti	Frazione
LET_IS	Esito finale (con causa)	0: Alive	84.06%
		1: Cardiogenic shock	6.47%
		2: Pulmonary edema	1.06%
		3: Myocardial rupture	3.18%
		4: Progress of congestive heart failure	1.35%
		5: Thromboembolism	0.71%
		6: Asystole	1.59%
		7: Ventricular fibrillation	1.59%

La variabile LET_IS si può considerare come variabile binaria(0:Vivo, 1:Morto)

Nome	Descrizione	Valori Assunti	Frazione
LET_IS	Esito finale	0: Alive	84.06%
		1: Dead	17,75%

Occorre notare che le complicanze non sono esclusive ma ogni esempio potrebbe presentare una o più complicazioni diverse:

FI	PR	JE	FI	A_	OT	RA	DR	ZS	RE	P_	LE
BR	ED	LU	BR	V_	EK	ZR	ES	N	C_	IM	T_
P	S	D_	_J	BL	_L	IV	SL		IM	_S	IS
RE	TA	TA	EL	OK	AN		ER			TE	
DS	H	H	UD							N	

663	
102	X
104	X
1	X X
35	X
19	X X
6	X X
192	X
10	X X
14	X X
11	X X
5	X X X
7	X X X

40									X				
1								X				X	
14								X	X				
3								X	X	X			
36						X							X
1						X				X			X
1						X			X				X
38					X								
21					X								X
1					X							X	
12					X					X			
1					X					X			X
1					X					X	X		
26					X				X				
5					X				X				X
2					X				X			X	
9					X				X	X			
4					X				X	X			X
2					X			X					
1					X			X	X	X			
1					X			X	X	X			X
1					X	X				X			X
1					X	X			X				X
17				X									
5				X									X
2				X						X			
1				X						X			X
7				X					X				

Cases	FI BR _P RE DS	PR ED S_ TA H	JE LU D_ TA H	FI BR _J EL UD	A_ V_ BL OK	OT EK _L AN	RA ZR IV	DR ES SL ER	ZS N	RE C_ IM	P_ IM _S TE N	LE T_ IS
-------	----------------------------	---------------------------	---------------------------	----------------------------	----------------------	----------------------	----------------	----------------------	---------	----------------	---------------------------	----------------

1					X			X				
3					X		X					X
1					X	X						
1					X	X						X
2					X	X				X		
1					X	X			X	X		
25				X								
8				X								X
1				X							X	
1				X						X		
1				X						X	X	
2				X					X			
1				X					X			X

MYOCARDIAL INFARCTION-COMPLICATIONS

1				X						X	X		X
3				X			X						X
1				X		X				X	X		
1				X		X		X			X		
1				X		X		X	X				
1				X	X								
1				X	X								X
2				X	X					X			
1				X	X		X						X
15			X										
1			X								X		X
5			X							X			
1			X							X	X		
1			X					X			X	X	
1			X			X							
1			X			X					X		X
1			X			X		X	X				
1			X		X								
1			X		X								X
1			X	X									X
1			X	X									X
1			X	X					X	X			
1			X	X				X			X		
1			X	X		X							
1			X	X	X					X			X
5		X											
1		X									X		
1		X									X	X	
2		X								X			
1		X				X					X		
1		X				X				X			
1		X	X										
1		X	X		X								
63		X											
5		X											X
3		X										X	
3		X									X		
5		X									X		X
1		X									X	X	
Cases	FI BR _P RE DS	PR ED S_ TA H	JE LU D_ TA H	FI BR _J EL UD	A_ V_ BL OK	OT EK _L AN	RA ZR IV	DR ES SL ER	ZS N	RE C_ IM	P_ IM _S TE N	LE T_ IS	
30	X								X				
2	X								X				X
1	X								X		X		

4	X								X	X		
1	X								X	X		X
1	X							X			X	
1	X							X	X			
1	X							X	X			X
1	X							X	X		X	
1	X							X				X
1	X							X		X		X
2	X							X	X			X
1	X							X	X			X
6	X					X						
1	X					X				X		
5	X					X			X			
1	X					X			X			X
1	X					X			X	X		X
2	X				X							
2	X				X				X			
1	X				X		X					X
1	X				X	X						X
3	X			X								
1	X			X								X
1	X			X						X		X
2	X			X					X			
2	X			X					X	X		
1	X			X		X						
1	X		X				X					X
1	X		X			X			X	X		
1	X		X		X							X
1	X		X	X						X		X
1	X		X	X				X	X			
1	X		X	X		X						
2	X	X										
2	X	X										X
1	X	X						X	X			
1	X	X				X			X			X
1	X	X		X								X

Cases

FI BR _P RE DS	PR ED S_ TA H	JE LU D_ TA H	FI BR _J EL UD	A_ V_ BL OK	OT EK _L AN	RA ZR IV	DR ES SL ER	ZS N	RE C_ IM	P_ IM _S TE N	LE T_ IS
----------------------------	---------------------------	---------------------------	----------------------------	----------------------	----------------------	----------------	----------------------	---------	----------------	---------------------------	----------------

4					X			X	X		X
1					X		X	X	X		
1					X	X			X		X
1					X	X		X			X
1				X	X			X	X		
1			X					X	X		X
1			X		X			X	X		
1			X		X		X		X		
1			X		X		X	X			

MYOCARDIAL INFARCTION-COMPLICATIONS

1			X	X		X				X
1		X					X		X	X
1		X			X				X	X
1		X			X		X	X		
1		X	X				X	X		
1		X	X				X	X		
1	X						X	X		X
1	X						X	X		X
1	X						X	X		X
1	X					X			X	X
2	X					X		X		X
1	X					X	X			X
1	X				X			X		X
1	X				X					X
1	X				X	X				X
1	X				X	X				X
2	X				X			X	X	
1	X		X			X				X
1	X		X		X					X
1	X		X	X		X				
1	X	X					X	X		
1	X	X			X					X
1					X		X	X	X	X
1		X	X	X				X		X
1	X				X			X	X	X
1	X	X			X			X	X	
1	X	X	X					X		X
1	X	X	X				X	X		
1	X	X			X			X		X

663 record non contengono complicazioni.

I record contengono fino a 5 complicazioni.

In totale 36 record hanno 4 complicazioni e 7 record contengono 5 complicazioni

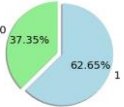
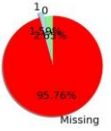
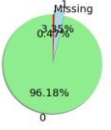
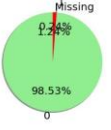
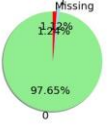
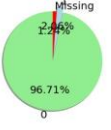
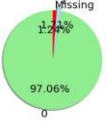
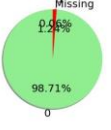
VARIABILI DI INPUT

QUANTITATIVE: 12



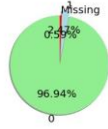

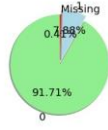
Name	Descrizione	Missing	Q1	Median	Q3	Min	Mean	Max	Std
AGE		8	54	63	70	26	61.857	92	11.2566
S_AD_KBRIG	Pressione sanguigna sistolica secondo il team di cardiologia d'emergenza	1076	120	140	160	0	136.907	260	34.9698
D_AD_KBRIG	Pressione sanguigna diastolica secondo il team di cardiologia d'emergenza	1076	70	80	90	0	81.3942	190	19.7292
S_AD_ORIT	Pressione sanguigna sistolica secondo l'unità di terapia intensiva	267	120	130	150	0	134.588	260	31.3374
D_AD_ORIT	Pressione sanguigna diastolica secondo l'unità di terapia intensiva	267	80	80	90	0	82.7495	190	18.3147
NA_BLOOD	Contenuto di sodio nel siero	375	133	136	140	117	136.551	169	6.50966
ALT_BLOOD	Contenuto di AIAT nel siero	284	0.23	0.38	0.61	0.03	0.481455	3	0.387124
AST_BLOOD	Contenuto di AsAT nel siero	285	0.15	0.22	0.33	0.04	0.263717	2.15	0.20173
KFK_BLOOD	Contenuto di CPK nel siero	1696	1.35	1.6	2.25	1.2	2	3.6	0.948683
L_BLOOD	Conteggio dei globuli bianchi	125	6.4	8	10.45	2	8.78291	27.9	3.39948
K_BLOOD	Contenuto di potassio nel siero	371	3.7	4.1	4.6	2.3	4.19	8.2	0.75
ROE	(Tasso di sedimentazione eritrocitaria)	203	5	10	18	1	13.4449	140	11.2925

CATEGORICHE BOOLEANE:78

MYOCARDIAL INFARCTION-COMPLICATIONS

Name	Description	1	o	Missings	Pie
SEX	Genere del pzazione o:Donna 1:Uomo	62.65%	37.35%	0.0%	
IBS_NASL	Ereditarietà su CHD. o:non presente 1 : presente	1.59%	2.65%	95.76%	
SIM_GIPERT	Presenza di ipertensione sintomatica	3.35%	96.18%	0.47%	
nr_11	Osservazione dell'aritmia nell'anamnesi	2.47%	96.29%	1.24%	
nr_01	Presenza di contrazioni atriali premature nell'anamnesi	0.24%	98.53%	1.24%	
nr_02	Contrazioni ventricolari premature nell'anamnesi	1.12%	97.65%	1.24%	
nr_03	Parossismi di fibrillazione atriale nell'anamnesi	2.06%	96.71%	1.24%	
nr_04	Presenza di una forma persistente di fibrillazione atriale nell'anamnesi	1.71%	97.06%	1.24%	
nr_07	Fibrillazione ventricolare nell'anamnesi	0.06%	98.71%	1.24%	
nr_08	Tachicardia parossistica ventricolare nell'anamnesi	0.24%	98.53%	1.24%	




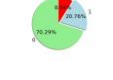



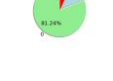






np_01	Blocco AV di primo grado nell'anamnesi	0.12%	98.82%	1.06%	
np_04	Blocco AV di terzo grado nell'anamnesi	0.18%	98.76%	1.06%	
np_05	LBBB (ramo anteriore) nell'anamnesi	0.65%	98.29%	1.06%	
np_07	LBBB incompleto nell'anamnesi	0.06%	98.88%	1.06%	
np_08	LBBB completo nell'anamnesi	0.35%	98.59%	1.06%	
np_09	RBBB incompleto nell'anamnesi	0.12%	98.82%	1.06%	

Name	Description	1	o	Missings	Pie
np_10	RBBB completo nell'anamnesi	0.18%	98.76%	1.06%	
endocr_01	Diabete mellito nell'anamnesi	13.41%	85.94%	0.65%	
endocr_02	L'obesità nell'anamnesi	2.47%	96.94%	0.59%	
endocr_03	Tireotossicosi nell'anamnesi	0.76%	98.65%	0.59%	
zab_leg_01	Bronchite cronica nell'anamnesi	7.88%	91.71%	0.41%	


MYOCARDIAL INFARCTION-COMPLICATIONS







zab_leg_o2	Bronchite cronica ostruttiva nell'anamnesi	7.12%	92.47%	0.41%	
zab_leg_o3	Asma bronchiale nell'anamnesi	2.18%	97.41%	0.41%	
zab_leg_o4	Asma bronchiale nell'anamnesi	0.53%	99.06%	0.41%	
zab_leg_o6	Tubercolosi polmonare nell'anamnesi	1.29%	98.29%	0.41%	
O_L_POST	Edema polmonare al momento dell'ammissione all'unità di terapia intensiva	6.47%	92.82%	0.71%	
K_SH_POST	Shock cardiogeno al momento dell'ammissione all'unità di terapia intensiva	2.71%	96.41%	0.88%	
MP_TP_POST	Parossismi di fibrillazione atriale al momento dell'ammissione all'unità di terapia intensiva unità di cura, (o in una fase preospedaliera)	6.71%	92.47%	0.82%	
SVT_POST	Parossismi di tachicardia sopraventricolare al momento dell'ammissione all'unità di terapia intensiva, (o in una fase pre-ospedaliera)	0.47%	98.82%	0.71%	
GT_POST	Parossismi di tachicardia ventricolare al momento del ricovero in unità di terapia intensiva, (o in una fase pre-ospedaliera)	0.47%	98.82%	0.71%	
FIB_G_POST	Fibrillazione ventricolare al momento dell'ammissione all'unità di terapia intensiva (o in una fase pre-ospedaliera)	0.88%	98.41%	0.71%	
IM_PG_P	Presenza di un infarto miocardico del ventricolo destro	2.94%	97.0%	0.06%	


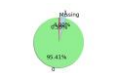

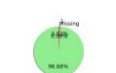

Name	Description	1	o	Missings	Pie
ritm_ecg_p_01	Ritmo ECG al momento dell'ammissione in ospedale - sinusale (con una frequenza cardiaca 60-90)	60.53%	30.53%	8.94%	

ritm_ecg_p_02	Ritmo ECG al momento dell'ammissione in ospedale fibrillazione atriale	5.59%	85.47%	8.94%	
ritm_ecg_p_04	Ritmo ECG al momento dell'ammissione in ospedale - atriale	1.35%	89.71%	8.94%	
ritm_ecg_p_06	Ritmo ECG al momento dell'ammissione in ospedale idioventricolare	0.06%	91.0%	8.94%	
ritm_ecg_p_07	Ritmo ECG al momento dell'ammissione in ospedale - seno con una frequenza cardiaca superiore a 90 (tachicardia)	20.76%	70.29%	8.94%	
ritm_ecg_p_08	Ritmo ECG al momento dell'ammissione in ospedale - seno con una frequenza cardiaca inferiore a 60 (bradicardia)	2.71%	88.35%	8.94%	
n_r_ecg_p_01	Contrazioni atriali premature su ECG al momento del ricovero in ospedale	3.41%	89.82%	6.76%	
n_r_ecg_p_02	Frequenti contrazioni atriali premature su ECG al momento del ricovero in ospedale	0.47%	92.76%	6.76%	
n_r_ecg_p_03	Contrazioni ventricolari premature su ECG al momento dell'ammissione in ospedale	12.0%	81.24%	6.76%	
n_r_ecg_p_04	Frequenti contrazioni ventricolari premature sull'ECG al momento dell'ammissione in ospedale	4.06%	89.18%	6.76%	
n_r_ecg_p_05	Parossismi di fibrillazione atriale su ECG al momento del ricovero in ospedale	4.12%	89.12%	6.76%	
n_r_ecg_p_06	Forma persistente di fibrillazione atriale su ECG al momento dell'ammissione in ospedale	1.88%	91.35%	6.76%	
n_r_ecg_p_08	Parossismi di tachicardia sopraventricolare su ECG al momento dell'ammissione in ospedale	0.24%	93.0%	6.76%	
n_r_ecg_p_09	Parossismi di tachicardia ventricolare su ECG al momento del ricovero in ospedale	0.12%	93.12%	6.76%	
n_r_ecg_p_10	Fibrillazione ventricolare su ECG al momento dell'ammissione in ospedale	0.12%	93.12%	6.76%	

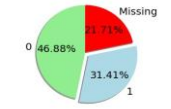

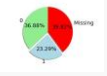



MYOCARDIAL INFARCTION-COMPLICATIONS


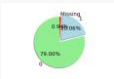
n_p_ecg_p_01	Blocco sinoatriale su ECG al momento dell'ammissione in ospedale	0.12%	93.12%	6.76%	
--------------	--	-------	--------	-------	---

Name	Description	1	o	Missings	Pie
n_p_ecg_p_03	Blocco AV di primo grado su ECG al momento dell'ammissione in ospedale	1.88%	91.35%	6.76%	
n_p_ecg_p_04	Blocco AV di secondo grado di tipo 1 (Mobitz I/Wenckebach) su ECG al al momento del ricovero in ospedale	0.29%	92.94%	6.76%	
n_p_ecg_p_05	Blocco AV di secondo grado di tipo 2 (Mobitz II/Hay) su ECG al momento dell'ammissione all'ospedale	0.12%	93.12%	6.76%	
n_p_ecg_p_06	Blocco AV di terzo grado su ECG al momento dell'ammissione in ospedale	1.59%	91.65%	6.76%	
n_p_ecg_p_07	LBBB (branca anteriore) sull'ECG al momento dell'ammissione in ospedale	6.0%	87.24%	6.76%	
n_p_ecg_p_08	LBBB (branca posteriore) su ECG al momento del ricovero in ospedale	0.41%	92.82%	6.76%	
n_p_ecg_p_09	LBBB incompleto su ECG al momento del ricovero in ospedale	0.59%	92.65%	6.76%	
n_p_ecg_p_10	LBBB completo su ECG al momento dell'ammissione in ospedale	2.0%	91.24%	6.76%	
n_p_ecg_p_11	RBBB incompleto su ECG al momento dell'ammissione in ospedale	1.65%	91.59%	6.76%	
n_p_ecg_p_12	RBBB completo su ECG al momento dell'ammissione in ospedale	4.59%	88.65%	6.76%	
fibr_ter_01	Terapia fibrinolitica con Celiasum 750k IU	0.76%	98.65%	0.59%	

fibr_ter_02	Terapia fibrinolítica con Celasum 1m IU	0.94%	98.47%	0.59%	
fibr_ter_03	Terapia fibrinolítica con Celasum 3m IU	4.0%	95.41%	0.59%	
fibr_ter_05	Terapia fibrinolítica con Streptase	0.24%	99.18%	0.59%	
fibr_ter_06	Terapia fibrinolítica con Celasum 500k IU	0.53%	98.88%	0.59%	
fibr_ter_07	Terapia fibrinolítica con Celasum 250k IU	0.35%	99.06%	0.59%	

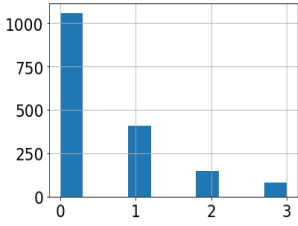
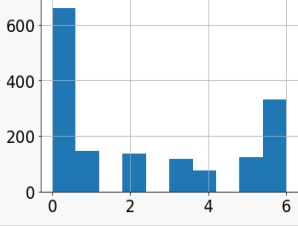
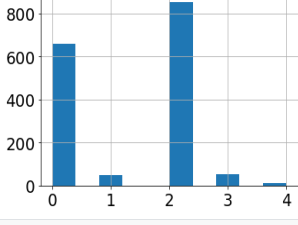
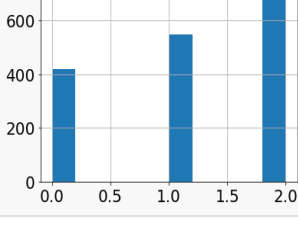
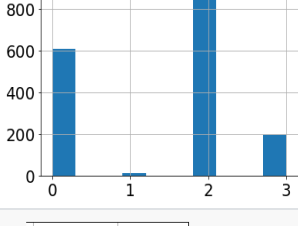
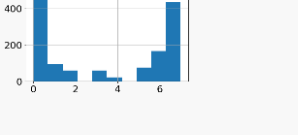
MYOCARDIAL INFARCTION-COMPLICATIONS

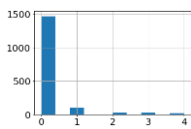
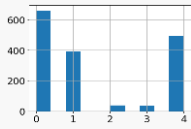
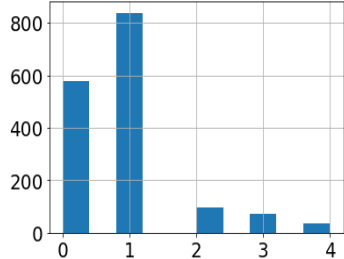
Name	Description	1	o	Missings	Pie
fibr_ter_o8	Terapia fibrinolitica con Streptodecase 1,5m IU	0.12%	99.29%	0.59%	
GIPO_K	Ipototassiemia (< 4 mmol/L)	31.41%	46.88%	21.71%	
GIPER_NA	Aumento del sodio nel siero	1.76%	76.18%	22.06%	
NA_KB	Uso di farmaci oppioidi da parte del team di cardiologia d'emergenza	36.35%	25.0%	38.65%	
NOT_NA_KB	Uso dei NSAID da parte del team di cardiologia d'urgenza	41.24%	18.41%	40.35%	
LID_KB	Uso della lidocaina da parte del team di cardiologia d'emergenza	23.29%	36.88%	39.82%	
NITR_S	Uso di nitrati liquidi in terapia intensiva	11.47%	88.0%	0.53%	
LID_S_n	Uso della lidocaina in terapia intensiva	28.18%	71.24%	0.59%	
B_BLOK_S_n	Uso dei beta-bloccanti in terapia intensiva	12.65%	86.71%	0.65%	
ANT_CA_S_n	Uso di calcio-antagonisti in terapia intensiva	66.18%	33.06%	0.76%	
GEPAR_S_n	Uso di a anticoagulanti (eparina) in terapia intensiva	70.76%	28.24%	1.0%	
ASP_S_n	Uso dell'acido acetilsalilico in terapia intensiva	73.65%	25.35%	1.0%	

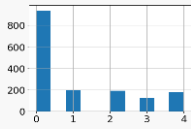
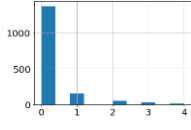
TIKL_S_n	Uso di Ticlid in terapia intensiva	1.76%	97.29%	0.94%	
TRENT_S_n	Uso di Trental in terapia intensiva	20.06%	79.0%	0.94%	

ORDINALI

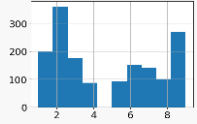
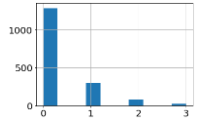
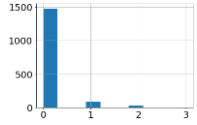
MYOCARDIAL INFARCTION-COMPLICATIONS

Name	Description	Values	Hist
INF_ANAM	Quantità di infarti miocardici nell'anamnesi	0,1,2,3 (missing: 4)	
STENOK_AN	Angina pectoris da sforzo nell'anamnesi	0-6 (missing:106)	
FK_STENOK	Classe funzionale (FC) dell'angina pectoris nell'ultimo anno	0-nessun angina pectoris, 1-I FC, 2- II FC, 3-III FC, 4-IV FC (missing : 73)	
IBS_POST	Malattia coronarica (CHD) nelle ultime settimane, giorni prima ricovero in ospedale	0-Nessun CHD, 1 -angina pectoris da sforzo, 2- angina pectoris instabile (missing :51)	
GB	Presenza di un'ipertensione essenziale	0-nessuna ipertensione essenziale, 1- Stadio 1, 2-Stadio 2, 3- Stadio 3 (missing : 9)	
DLIT_AG	Durata dell'ipertensione arteriosa	0- nessun ipertensione arteriosa, 1,2,4,5 (anni), 6 (da 6-10 anni), 7 (più di 10 anni) (missings : 248)	

ZSN_A	<p>Presenza di insufficienza cardiaca cronica (HF) nell'anamnesi, Attributo parzialmente ordinato: ci sono due linee di gravità: 0<1<2<4, 0<1<3<4. Lo stato 4 significa gli stati 2 e 3 simultanei</p>	<p>0 : - non c'è insufficienza cardiaca cronica 1: Stadio I 2:IIA stadio(insufficienza cardiaca dovuta a disfunzione sistolica del ventricolo destro) Stadio IIA (insufficienza cardiaca dovuta a disfunzione sistolica ventricolare sinistra) 4:Stadio IIB (insufficienza cardiaca dovuta a disfunzione sistolica del ventricolo destro e sinistro disfunzione) (missing: 54)</p>	 <table><caption>Data for ZSN_A Histogram</caption><thead><tr><th>Category</th><th>Count</th></tr></thead><tbody><tr><td>0</td><td>1400</td></tr><tr><td>1</td><td>100</td></tr><tr><td>2</td><td>50</td></tr><tr><td>3</td><td>20</td></tr><tr><td>4</td><td>10</td></tr></tbody></table>	Category	Count	0	1400	1	100	2	50	3	20	4	10
Category	Count														
0	1400														
1	100														
2	50														
3	20														
4	10														
ant_im	<p>Presenza di un infarto miocardico anteriore (ventricolare sinistro) (cambiamenti ECG nelle derivazioni V1 - V4)</p>	<p>0: Nessuna presenza 1:Il QRS non ha cambiamenti 2:- Il QRS è come il complesso QR 3:- Il QRS è come il complesso Qe 4- Il QRS è come il complesso QS (missings: 83)</p>	 <table><caption>Data for ant_im Histogram</caption><thead><tr><th>Category</th><th>Count</th></tr></thead><tbody><tr><td>0</td><td>600</td></tr><tr><td>1</td><td>400</td></tr><tr><td>2</td><td>50</td></tr><tr><td>3</td><td>20</td></tr><tr><td>4</td><td>500</td></tr></tbody></table>	Category	Count	0	600	1	400	2	50	3	20	4	500
Category	Count														
0	600														
1	400														
2	50														
3	20														
4	500														
lat_im	<p>Presenza di un infarto miocardico laterale (ventricolare sinistro) (cambiamenti ECG nelle derivazioni V5 - V6, I, AVL)</p>	<p>0: Nessuna presenza 1:Il QRS non ha cambiamenti 2:- Il QRS è come il complesso QR 3:- Il QRS è come il complesso Qe 4- Il QRS è come il complesso QS (missings: 80)</p>	 <table><caption>Data for lat_im Histogram</caption><thead><tr><th>Category</th><th>Count</th></tr></thead><tbody><tr><td>0</td><td>600</td></tr><tr><td>1</td><td>800</td></tr><tr><td>2</td><td>100</td></tr><tr><td>3</td><td>80</td></tr><tr><td>4</td><td>50</td></tr></tbody></table>	Category	Count	0	600	1	800	2	100	3	80	4	50
Category	Count														
0	600														
1	800														
2	100														
3	80														
4	50														

Name	Description	Values	Hist												
inf_im	Presenza di un infarto miocardico inferiore (ventricolare sinistro) (cambiamenti ECG nelle derivazioni III, AVF, II)	0: Nessuna presenza 1:Il QRS non ha cambiamenti 2:- Il QRS è come il complesso QR 3:- Il QRS è come il complesso Qe 4- Il QRS è come il complesso QS (missings: 80)	 <table><caption>Data for inf_im Histogram</caption><thead><tr><th>Category</th><th>Count</th></tr></thead><tbody><tr><td>0</td><td>800</td></tr><tr><td>1</td><td>200</td></tr><tr><td>2</td><td>200</td></tr><tr><td>3</td><td>100</td></tr><tr><td>4</td><td>200</td></tr></tbody></table>	Category	Count	0	800	1	200	2	200	3	100	4	200
Category	Count														
0	800														
1	200														
2	200														
3	100														
4	200														
post_im	Presenza di un infarto miocardico posteriore (ventricolare sinistro) (cambiamenti ECG in V7 - V9, cambiamenti di reciprocità nelle derivazioni V1 - V3)	0: Nessuna presenza 1:Il QRS non ha cambiamenti 2:- Il QRS è come il complesso QR 3:- Il QRS è come il complesso Qe 4- Il QRS è come il complesso QS (missings: 72)	 <table><caption>Data for post_im Histogram</caption><thead><tr><th>Category</th><th>Count</th></tr></thead><tbody><tr><td>0</td><td>1100</td></tr><tr><td>1</td><td>100</td></tr><tr><td>2</td><td>50</td></tr><tr><td>3</td><td>20</td></tr><tr><td>4</td><td>10</td></tr></tbody></table>	Category	Count	0	1100	1	100	2	50	3	20	4	10
Category	Count														
0	1100														
1	100														
2	50														
3	20														
4	10														

MYOCARDIAL INFARCTION-COMPLICATIONS

TIME_B_S	Tempo trascorso dall'inizio dell'attacco di CHD al ospedale	1: meno di 2 ore 2: 2-4 ore 3: 4-6 ore 4: 6-8 ore 5 : 8-12 ore 6 : 12-24 ore 7 : più di un giorno 8: Più di 2 giorni 9: oltre 3 giorni (missings: 126)	
R_AB_1_n	Ricaduta del dolore nelle prime ore del periodo di ricovero	0: Nessuna ricaduta 1:solo una 2: 2 volte 3: 3 o più volte (missings : 16)	
R_AB_2_n	Ricaduta del dolore nel secondo giorno del periodo di ricovero	0: Nessuna ricaduta 1:solo una 2: 2 volte 3: 3 o più volte (missings: 108)	
R_AB_3_n	Ricaduta del dolore nel terzo giorno del periodo ospedaliero	0: Nessuna ricaduta 1:solo una 2: 2 volte 3: 3 o più volte (missings: 128)	
NA_R_1_n	Uso di farmaci oppioidi in terapia intensiva nelle prime ore del periodo	0 : No 1: Una volta 2: 2 Volte 3: 3volte 4: 4 volte (missings: 5)	
NA_R_2_n	Uso di farmaci oppioidi in terapia intensiva nel secondo giorno di ricovero periodo	0 : No 1: Una volta 2: 2 Volte 3: 3volte (missings:108)	
NA_R_3_n	Uso di farmaci oppioidi in terapia intensiva nel terzo giorno di ricovero periodo	0 : No 1: Una volta 2: 2 Volte 3: 3volte (missings: 131)	
Name	Description	Values	Hist
NOT_NA_1_n	Uso di NSAID in terapia intensiva nelle prime ore di degenza	0 : No 1: Una volta 2: 2 Volte 3: 3volte 4: 4 volte (missings: 10)	

NOT_NA_2_n	Uso di NSAID in terapia intensiva nel secondo giorno di degenza	0 : No 1: Una volta 2: 2 Volte 3: 3volte 4: 4 volte (missings: 110)	
NOT_NA_3_n	Uso di NSAID in terapia intensiva nel terzo giorno di degenza	0 : No 1: Una volta 2: 2 Volte (missings: 131)	

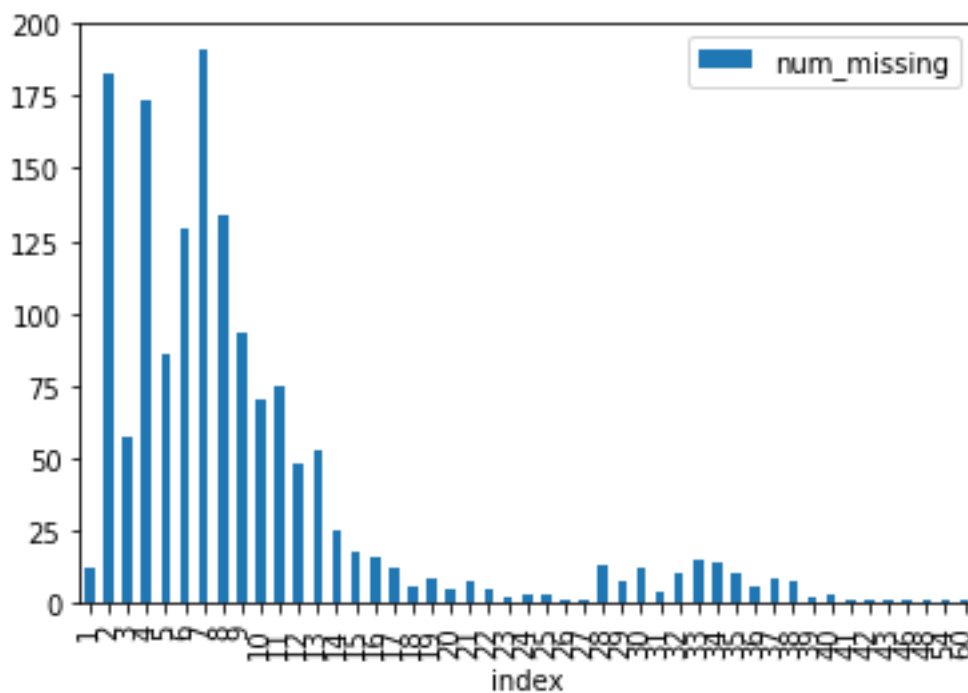
DATA PREPARATION

PROBLEMI RISCONTRATI

- **Valori Mancanti**
- È un **dataset multi label** (12 variabili di output), occorre definire un metodo di gestione delle etichette
- Dataset **sbilanciato**

MISSING DATA HISTOGRAM

Dato l'elevato numero di variabile un modo efficace per visualizzare quanto più globalmente possibile la situazioni di valori mancanti tra le osservazioni è il seguente istogramma:



MYOCARDIAL INFARCTION-COMPLICATIONS

Nota bene: Va letto come " ci sono meno di 25 osservazioni con 1 valore mancante (index 1), 200 osservazioni con 2 valori mancanti (index 2) e così via..."

LISTA DI PERCENTUALE DI DATI MANCANTI

ID - 0%	GB - 1%	np_07 - 1%
SEX - 0%	fibr_ter_01 - 1%	np_08 - 1%
FIBR_PREDS - 0%	fibr_ter_02 - 1%	np_09 - 1%
PREDS_TAH - 0%	fibr_ter_03 - 1%	np_10 - 1%
JELUD_TAH - 0%	fibr_ter_05 - 1%	GEPAR_S_n - 1%
FIBR_JELUD - 0%	fibr_ter_06 - 1%	ASP_S_n - 1%
A_V_BLOK - 0%	fibr_ter_07 - 1%	nr_11 - 1%
OTEK_LANC - 0%	fibr_ter_08 - 1%	nr_01 - 1%
RAZRIV - 0%	endocr_01 - 1%	nr_02 - 1%
DRESSLER - 0%	endocr_02 - 1%	nr_03 - 1%
ZSN - 0%	endocr_03 - 1%	nr_04 - 1%
REC_IM - 0%	O_L_POST - 1%	nr_07 - 1%
P_IM_STEN - 0%	SVT_POST - 1%	nr_08 - 1%
LET_IS - 0%	GT_POST - 1%	IBS_POST - 2%
IM_PG_P - 0%	FIB_G_POST - 1%	ZSN_A - 3%
INF_ANAM - 0%	LID_S_n - 1%	FK_STENOK - 4%
NA_R_1_n - 0%	B_BLOK_S_n - 1%	post_im - 4%
zab_leg_01 - 0%	MP_TP_POST - 1%	lat_im - 4%
zab_leg_02 - 0%	ANT_CA_S_n - 1%	ant_im - 5%
zab_leg_03 - 0%	K_SH_POST - 1%	inf_im - 5%
zab_leg_04 - 0%	R_AB_1_n - 1%	STENOK_AN - 6%
zab_leg_06 - 0%	TIKL_S_n - 1%	R_AB_2_n - 6%
AGE - 1%	TRENT_S_n - 1%	NA_R_2_n - 6%
SIM_GIPERT - 1%	np_01 - 1%	NOT_NA_2_n - 6%
NITR_S - 1%	np_04 - 1%	n_p_ecg_p_01 - 7%
NOT_NA_1_n - 1%	np_05 - 1%	n_p_ecg_p_03 - 7%

n_p_ecg_p_04 - 7%	NA_R_3_n - 7%	NA_BLOOD - 22%
n_p_ecg_p_05 - 7%	NOT_NA_3_n - 7%	NA_KB - 39%
n_p_ecg_p_06 - 7%	TIME_B_S - 7%	LID_KB - 40%
n_p_ecg_p_07 - 7%	L_BLOOD - 8%	NOT_NA_KB - 41%
n_p_ecg_p_08 - 7%	ritm_ecg_p_01 - 9%	S_AD_KBRIG - 63%
n_p_ecg_p_09 - 7%	ritm_ecg_p_02 - 9%	D_AD_KBRIG - 63%
n_p_ecg_p_10 - 7%	ritm_ecg_p_04 - 9%	IBS_NASL - 96%
n_p_ecg_p_11 - 7%	ritm_ecg_p_06 - 9%	KFK_BLOOD - 100%
n_p_ecg_p_12 - 7%	ritm_ecg_p_07 - 9%	
n_r_ecg_p_01 - 7%	ritm_ecg_p_08 - 9%	
n_r_ecg_p_02 - 7%	ROE - 12%	
n_r_ecg_p_03 - 7%	DLIT_AG - 15%	
n_r_ecg_p_04 - 7%	S_AD_ORIT - 17%	
n_r_ecg_p_05 - 7%	D_AD_ORIT - 17%	
n_r_ecg_p_06 - 7%	ALT_BLOOD - 17%	
n_r_ecg_p_08 - 7%	AST_BLOOD - 17%	
n_r_ecg_p_09 - 7%	GIPO_K - 21%	
n_r_ecg_p_10 - 7%	K_BLOOD - 22%	
R_AB_3_n - 7%	GIPER_NA - 22%	

DATA CLEANING

Per risolvere il problema dei dati mancanti, si sono effettuate le seguenti procedure

DROP DI VARIABILI

Eliminazione di intere colonne di valori di variabili con percentuale di dati mancanti maggiori di una determinata soglia.

Si è scelto di eliminare le variabili che superano una soglia del **40%** di dati mancanti, quindi le variabili:

- **S_AD_KBRIG** (Pressione sanguigna sistolica secondo il team di cardiologia d'emergenza),
- **D_AD_KBRIG** (Pressione sanguigna diastolica secondo il team di cardiologia d'emergenza),
- **IBS_NASL** (Ereditarietà su CHD),
- **KFK_BLOOD** (Contenuto di CPK nel siero)
- **NOT_NA_KB** (Uso dei NSAID da parte del team di cardiologia d'urgenza)

MYOCARDIAL INFARCTION-COMPLICATIONS

DROP DI OSSERVAZIONI

Per eliminare le osservazioni senza troppe perdite, occorre stabilire un valore di soglia di valori mancanti oltre il quale si consideri la scelta di eliminare quell'osservazione, in modo da non eliminare troppe (o troppe poche) osservazioni.

Scegliendo un valore di soglia di **20** sono rimaste 1570 osservazioni.

SOSTITUZIONE DEL VALORE MANCANTE CON IL VALORE PIÙ PROBABILE

I valori mancanti rimanenti devono essere sostituiti in qualche modo.

Per prevedere il valore più probabile sulla base delle altre informazioni presenti nel dataset si è scelto di utilizzare il metodo **Nearest Neighbour Imputation**.

I valori mancanti di ogni campione sono imputati utilizzando il valore medio dei K vicini più vicini trovati nel dataset. Due osservazioni sono vicine se le features (con valori non mancanti) sono vicine,.

Si è stabilito un numero di vicini **K = 3**

DATA SCALING

Considerando solo le variabili numeriche del Dataset

('AGE','S_AD_ORIT','D_AD_ORIT','ALT_BLOOD','L_BLOOD','K_BLOOD','ROE')

Si normalizzano i valori delle variabili mediante una normalizzazione Z-score

$$Z_i = \frac{x - \mu_i}{\sigma_i}$$

FACTOR ANALYSIS

L'analisi dei fattori è una tecnica che viene utilizzata per ridurre un gran numero di variabili

in un minor numero di fattori. Questa tecnica estrae la massima varianza comune da tutte

le variabili e le mette in uno score comune.

La scelta del numero di fattori da considerare nella soluzione fattoriale può essere

effettuata secondo differenti criteri, in questo caso si è utilizzato il criterio di Kaiser, in base

al quale si considerano tutti i fattori il cui autovalore sia superiore o uguale a 1, il numero di

fattori scelto è 32.

MYOCARDIAL INFARCTION-COMPLICATIONS

MODELING

ANALISI DEI CLUSTER

L'obiettivo di Data Mining consiste in un task di clusterizzazione.

Occorre quindi determinare la tecnica di modellazione da utilizzare, quindi determinare l'**algoritmo di clustering**.

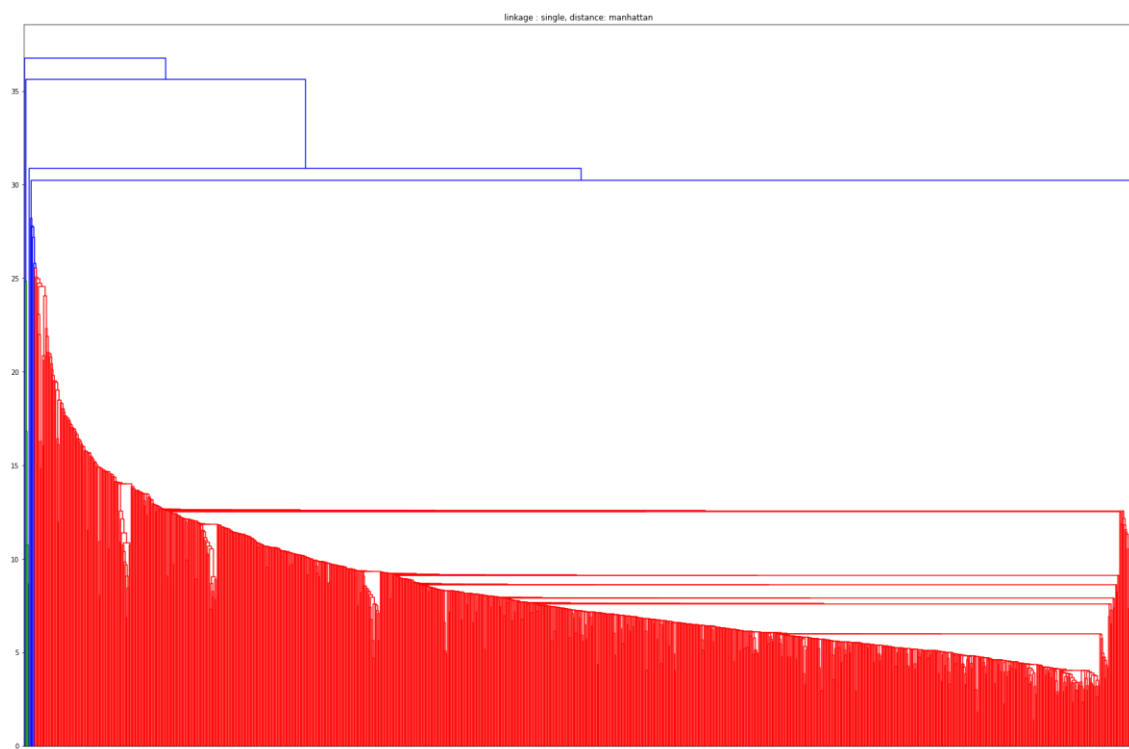
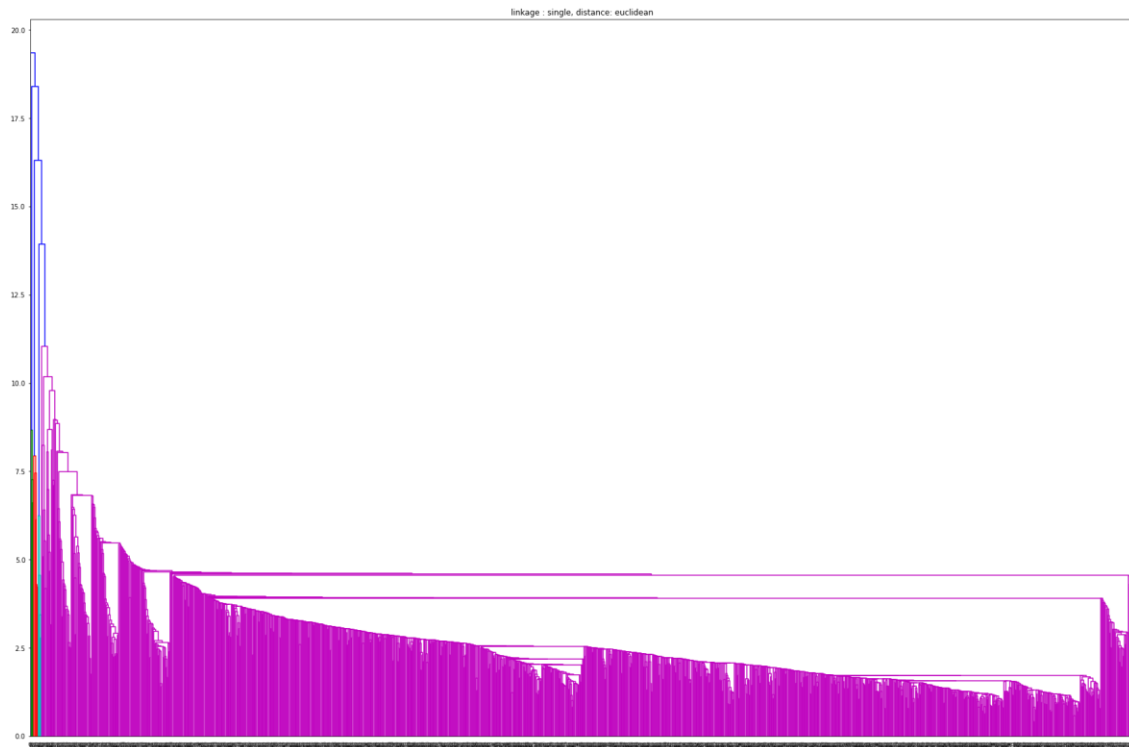
ALGORITMO DI CLUSTERING : MODELLO + PARAMETRI

Si considerano gli algoritmi di clustering gerarchico di tipo Agglomerativo, occorre effettuare una ricerca dei parametri ottimali:

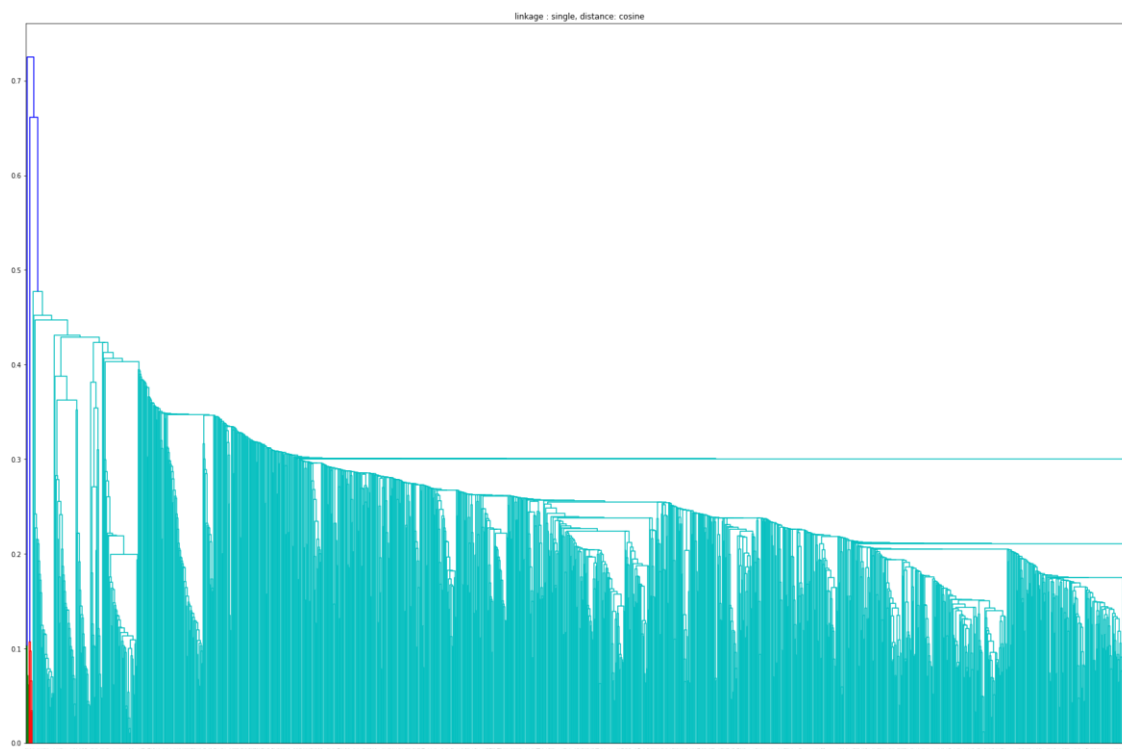
- Il parametro di k : **2 – n**
 - Il metodo di linkage : **Single, Average, Complete, Ward**
- La misura di distanza : **Euclidea, Coseno, Manhattan**

DENDOGRAMMI:

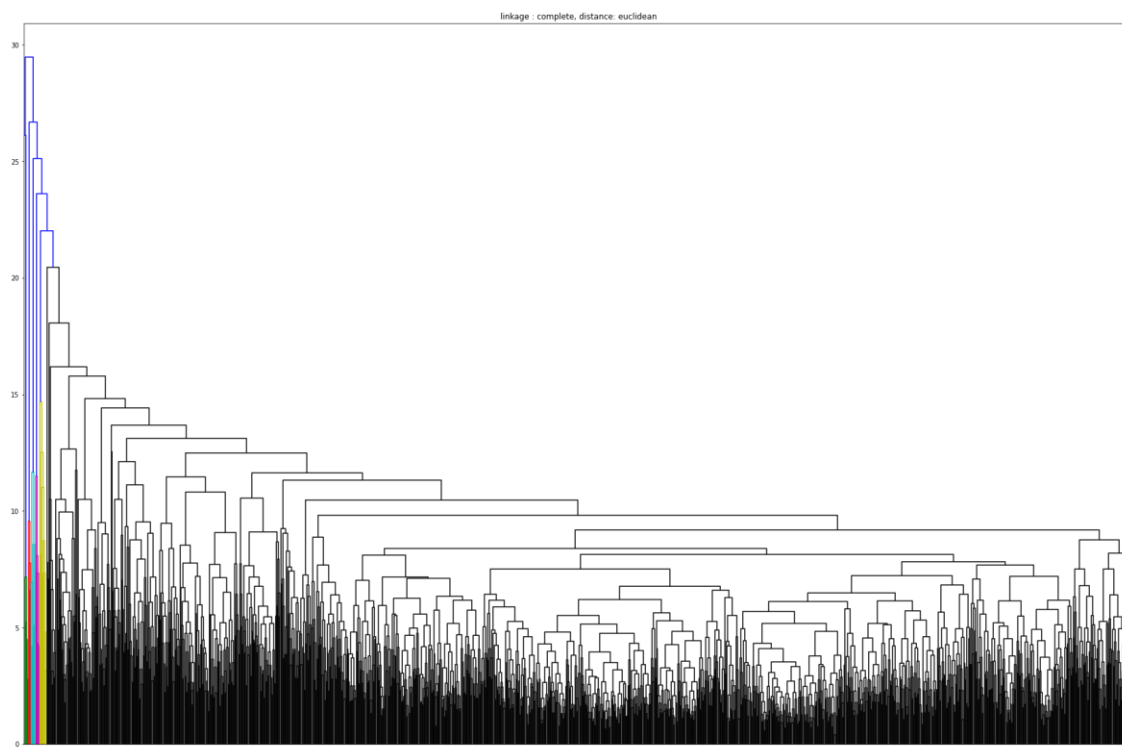
SINGLE LINKAGE

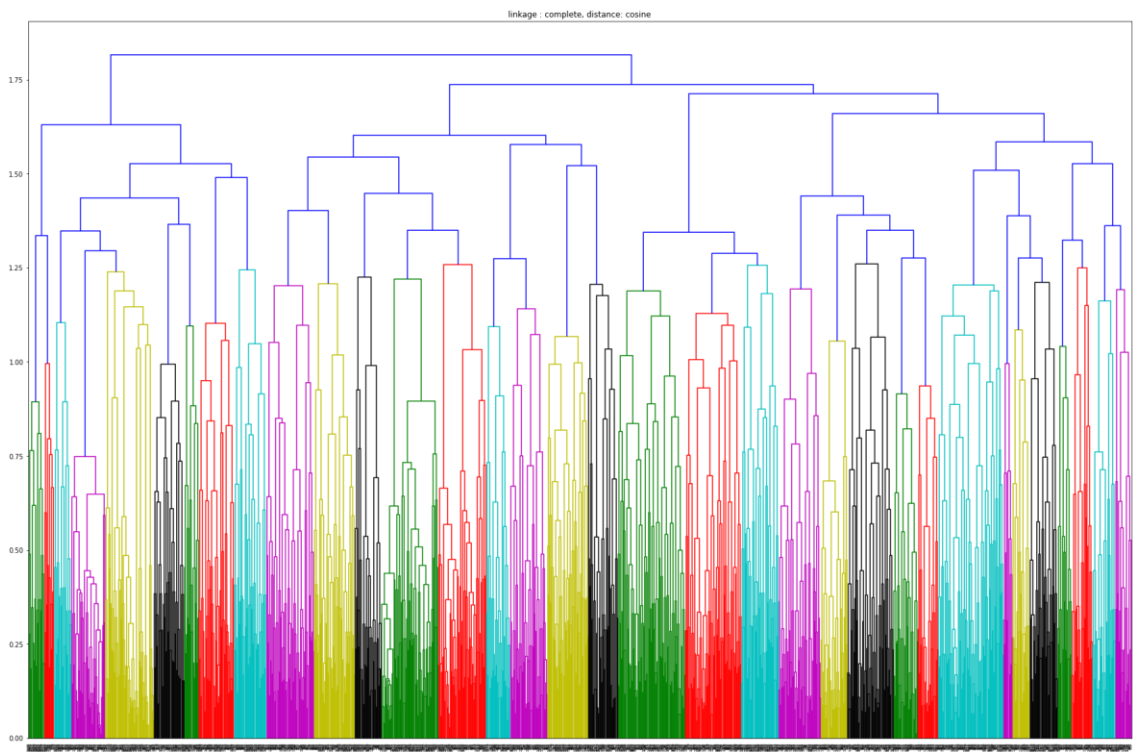
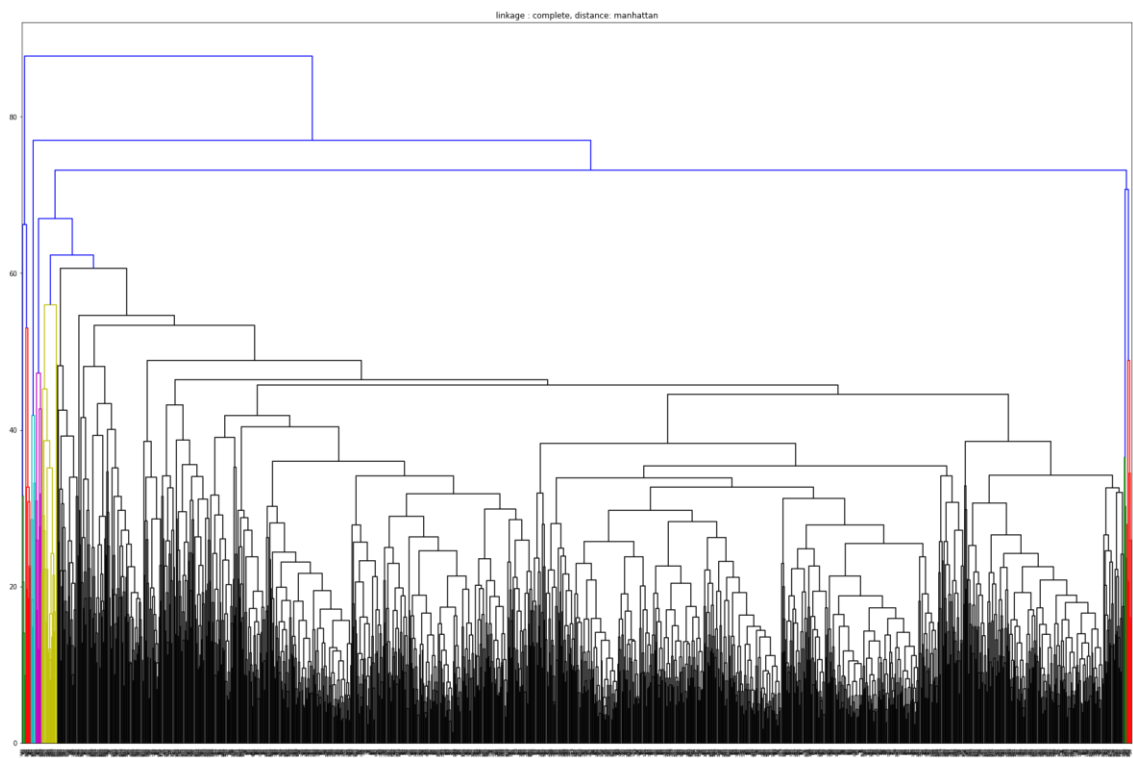


MYOCARDIAL INFARCTION-COMPLICATIONS



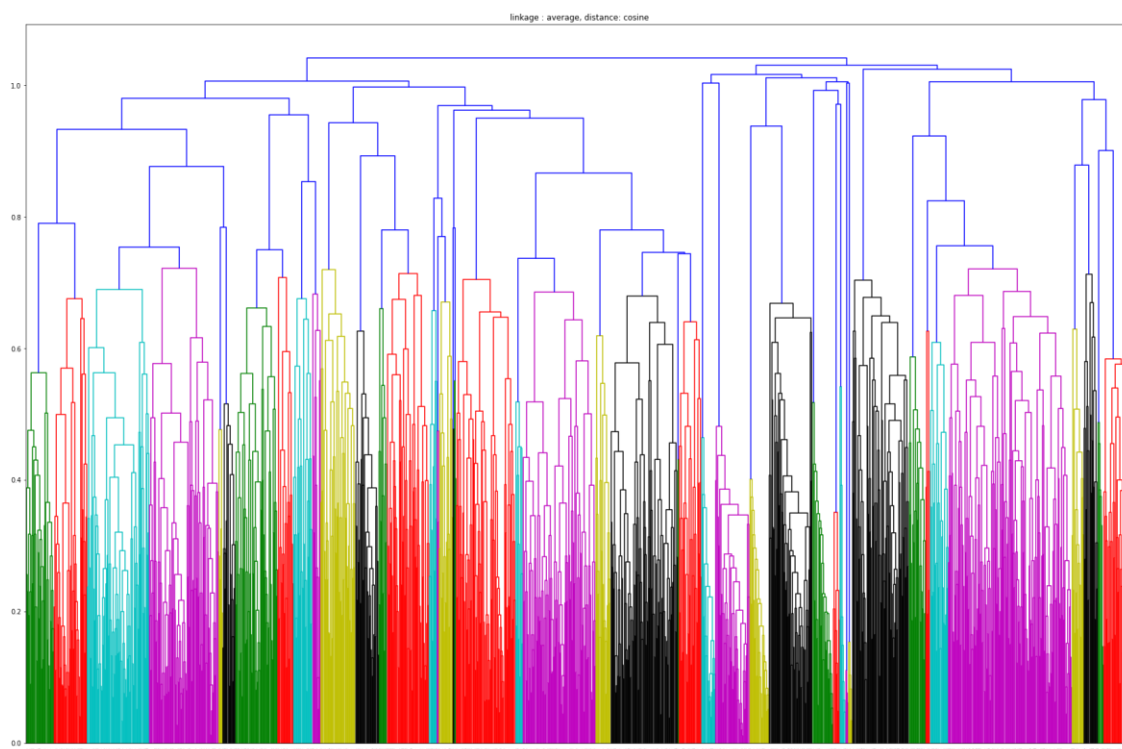
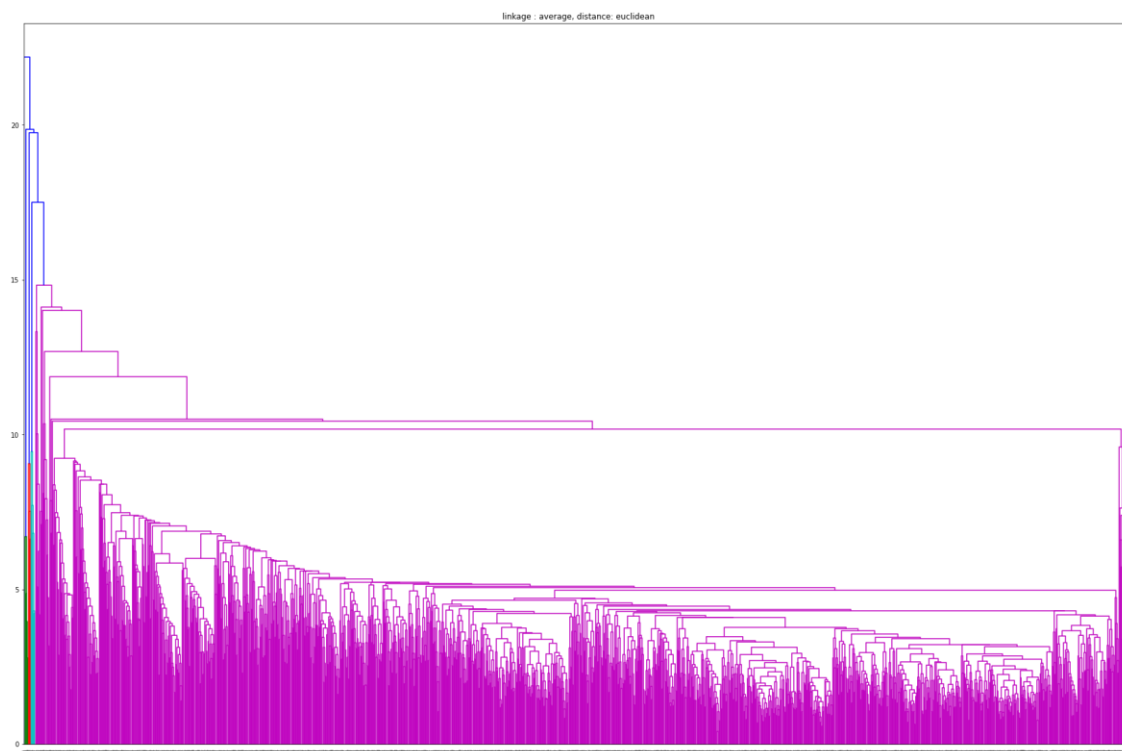
COMPLETE LINKAGE

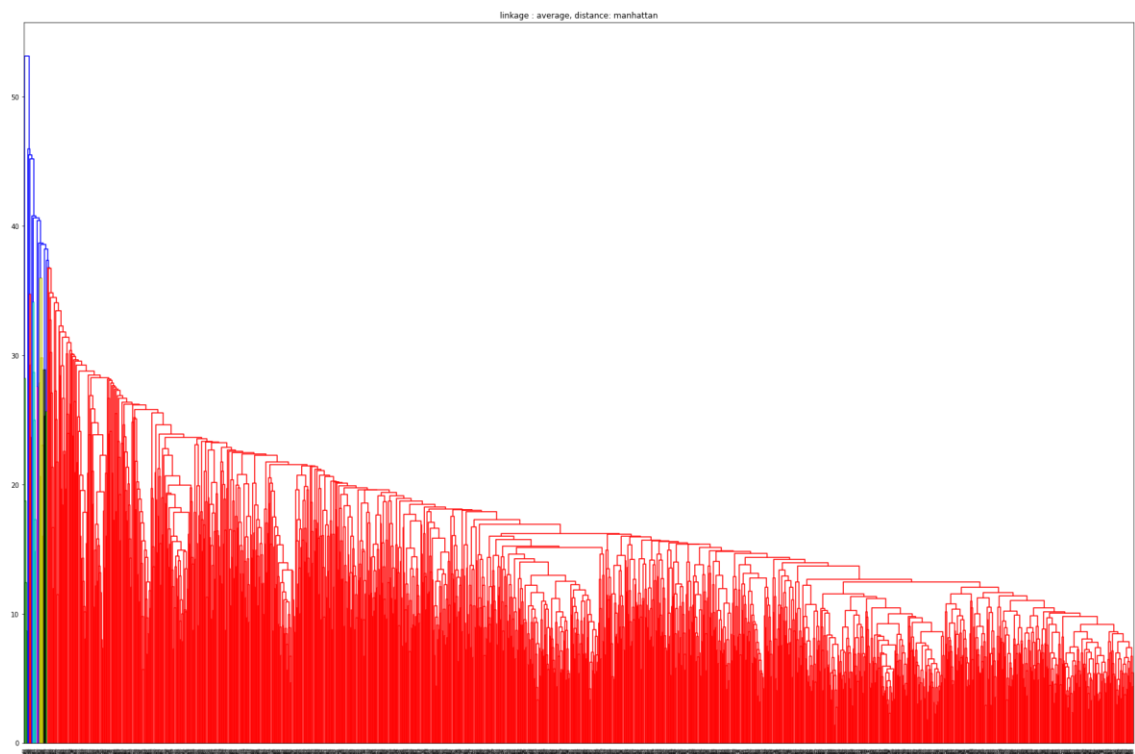




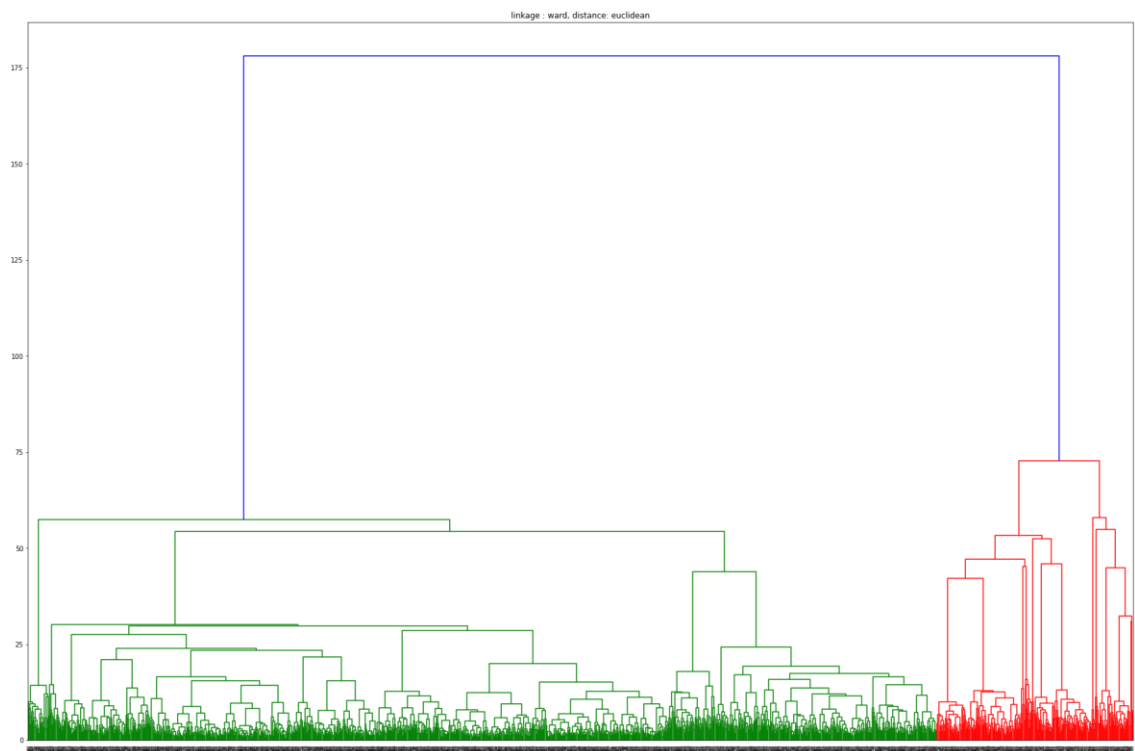
MYOCARDIAL INFARCTION-COMPLICATIONS

AVERAGE LINKAGE





WARD



MYOCARDIAL INFARCTION-COMPLICATIONS

CRITERI DI VALUTAZIONE:

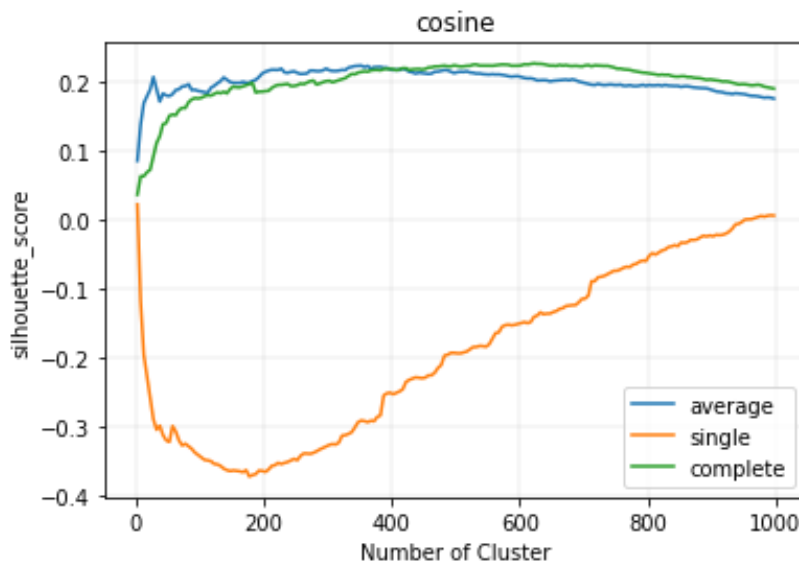
Due tipi di misure di convalida possono essere utilizzate per misurare le somiglianze tra le soluzioni di clustering: interne ed esterne. Il primo tipo di criterio misura gli attributi presi dai dati stessi e dai cluster formati, come la compattezza e la separabilità dei dati. Il secondo fa un confronto tra soluzioni di clustering, prendendone una come riferimento e confrontandola con altri raggruppamenti, in questo caso le label di output del dataset

CRITERI INTERI : INDICE DI SILHOUETTE

L'indice di **Silhouette** misura quanto un oggetto è simile al proprio cluster (coesione) rispetto agli altri cluster (separazione). La silhouette varia da -1 a +1, dove un valore alto indica che l'oggetto è ben assegnato al proprio cluster e poco assegnabile ai cluster vicini.

Se la maggior parte degli oggetti ha un valore alto, allora la soluzione di clustering è appropriata. Se molti punti hanno un valore basso o negativo, allora la configurazione di clustering potrebbe avere troppi o troppo pochi cluster.

COSINE



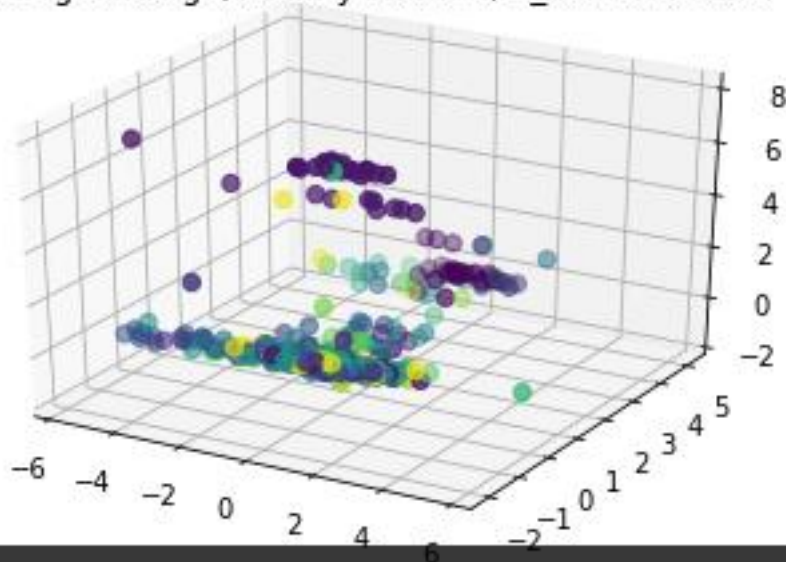
BUILD MODEL

MODEL OUTPUT

```
=====
averageLinkage, Affinity : cosine, n_clusters : 350
Silhouette : 0.22112826343378497
```

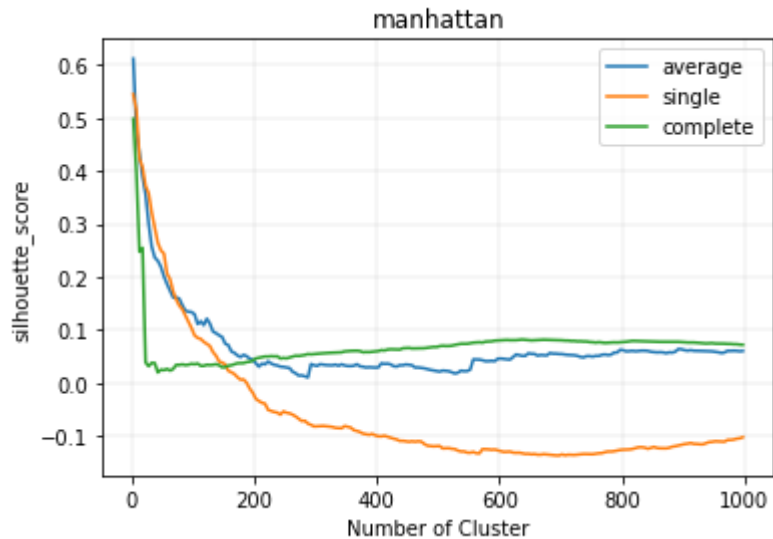
```
number of cluster of size 8 : 6
number of cluster of size 3 : 57
number of cluster of size 14 : 2
number of cluster of size 5 : 19
number of cluster of size 6 : 23
number of cluster of size 39 : 1
number of cluster of size 13 : 4
number of cluster of size 32 : 1
number of cluster of size 9 : 10
number of cluster of size 2 : 82
number of cluster of size 4 : 34
number of cluster of size 20 : 1
number of cluster of size 10 : 5
number of cluster of size 7 : 13
number of cluster of size 28 : 1
number of cluster of size 19 : 4
number of cluster of size 11 : 4
number of cluster of size 22 : 1
number of cluster of size 43 : 1
number of cluster of size 12 : 3
number of cluster of size 1 : 73
number of cluster of size 15 : 2
number of cluster of size 17 : 1
number of cluster of size 26 : 1
number of cluster of size 21 : 1
```

average-linkage, Affinity : cosine, n_clusters : 350



MYOCARDIAL INFARCTION-COMPLICATIONS

MANHATTAN

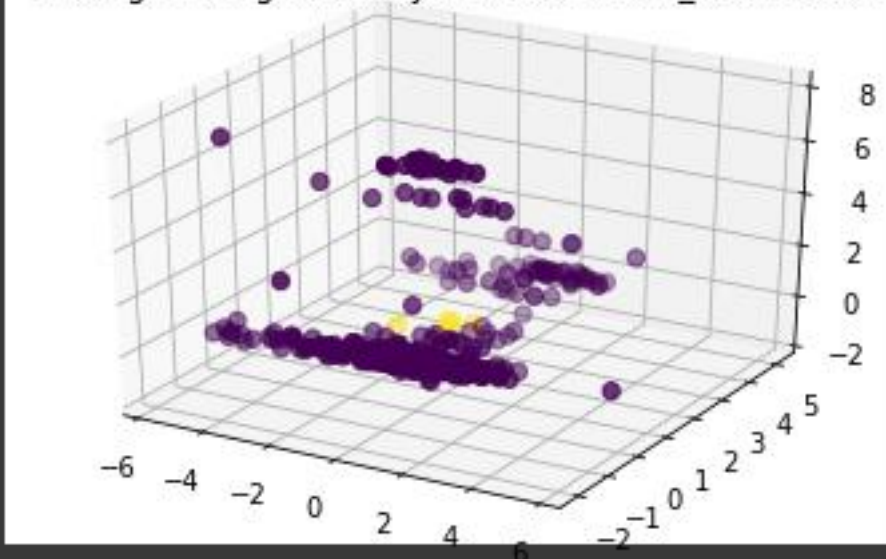


BUILD MODEL

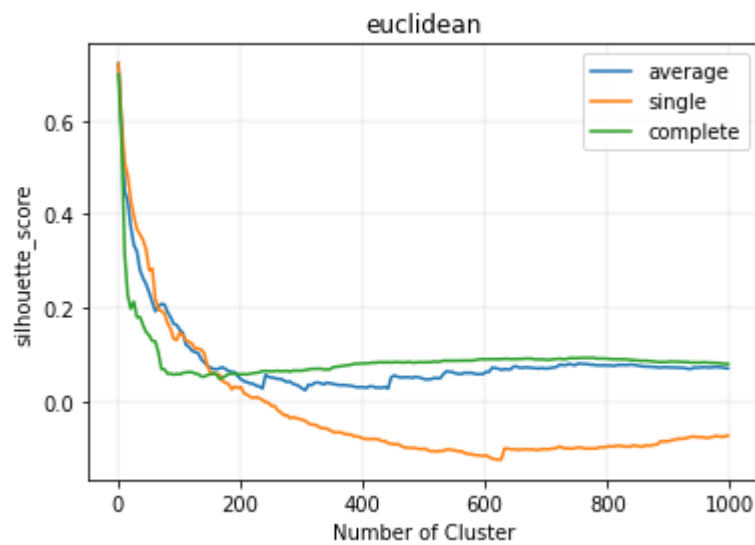
MODEL OUTPUT

```
=====
averageLinkage, Affinity : manhattan, n_clusters : 2
Silhouette : 0.6114956890733572
number of cluster of size 1565 : 1
number of cluster of size 5 : 1
```

average-linkage, Affinity : manhattan, n_clusters : 2



EUCLIDEAN

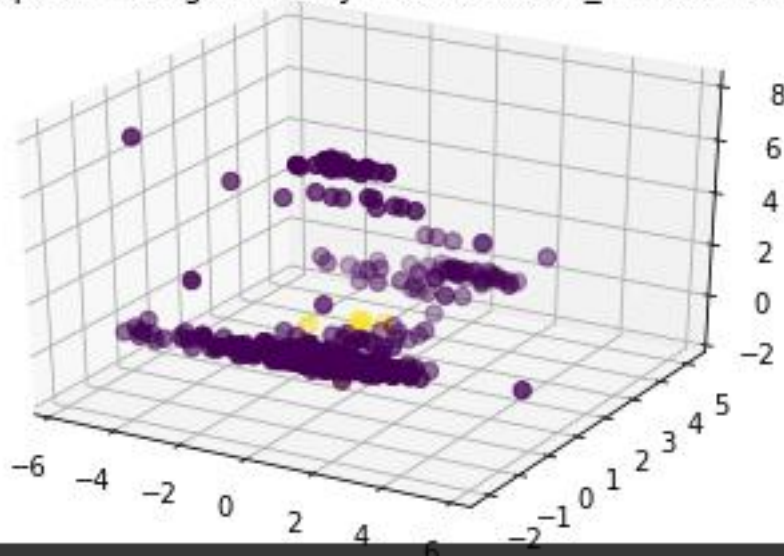


BUILD MODEL

MODEL OUTPUT

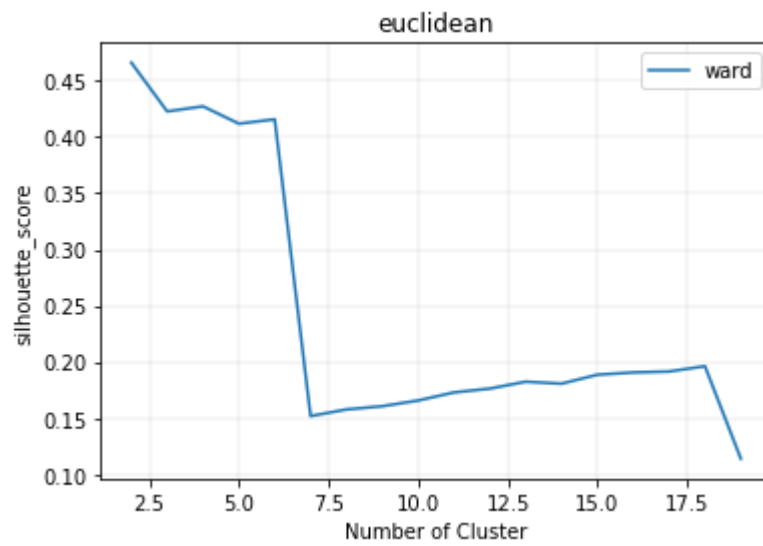
```
=====
completeLinkage, Affinity : euclidean, n_clusters : 2
Silhouette : 0.698423845297036
number of cluster of size 1564 : 1
number of cluster of size 6 : 1
```

complete-linkage, Affinity : euclidean, n_clusters : 2



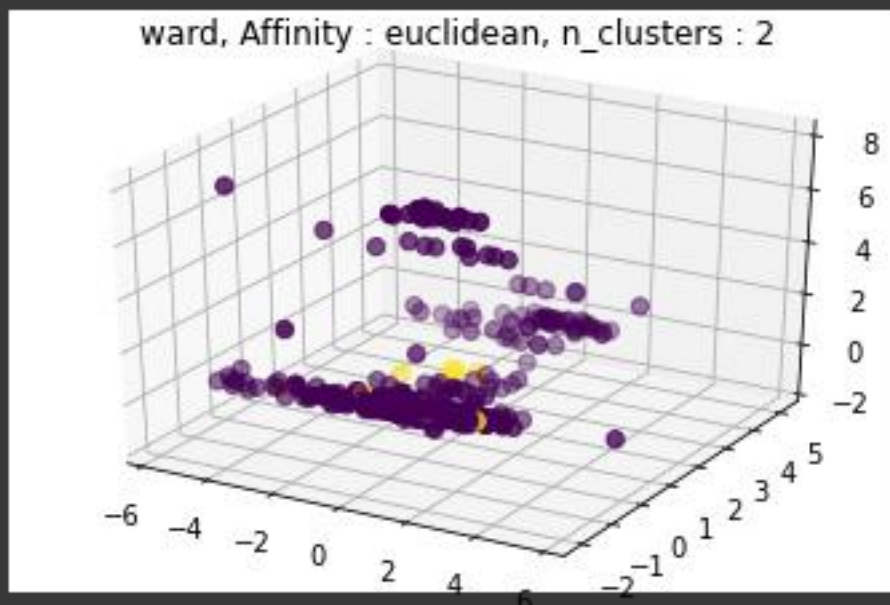
MYOCARDIAL INFARCTION-COMPLICATIONS

WARD + EUCLIDEAN



BUILD MODEL

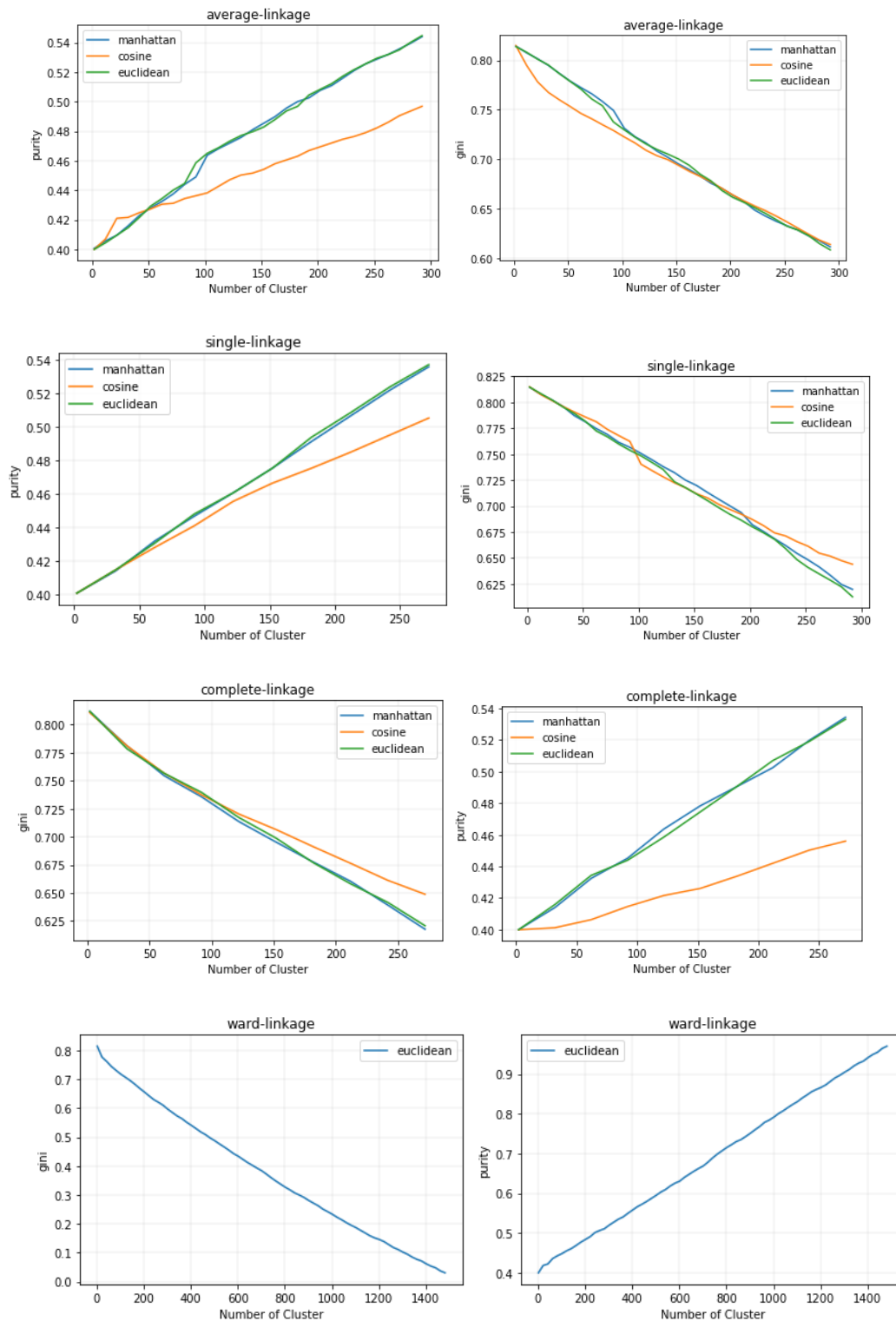
```
Silhouette : 0.491189944795383  
number of cluster of size 1522 : 1  
number of cluster of size 48 : 1
```



CRITERI ESTERNI

Per valutare i risultati del cluster scelto potrei utilizzare le variabili di output presenti nel dataset e un quindi un criterio di validazione esterna, tali criteri permettono di quantificare le somiglianze tra due soluzioni di clustering.

GINI E PURITY



MYOCARDIAL INFARCTION-COMPLICATIONS

Problema: I criteri esterni Gini e Purity tendono a valutare meglio soluzioni con un numero di cluster elevato, sino ad arrivare a tanti cluster quante sono le osservazioni.

Questo potrebbe essere dovuto al fatto che le **variabili di output a disposizione non sono mutuamente esclusive**, ma ogni osservazione è etichetta con più di una label.

Utilizzando un criterio di valutazione classici ogni singola possibile combinazione di label (complicazioni) verrebbe considerata come una classe arrivando a considerare un numero di classi pari a 133.

La matrice di confusione risultate sarebbe una matrice con un alto livello di confusione e ogni cluster conterrà elementi molto eterogenei loro

SOLUZIONE: INDICE OC

Con criteri esterni classici, è possibile solamente quantificare quanto sono simili due **soluzioni disgiunte**, dove ogni oggetto può essere etichettato con una sola label.

Utilizzare come criterio esterno una misura che consideri la probabilità che una qualsiasi coppia di oggetti possa essere trovata in una data soluzione o in entrambe le soluzioni di clustering, che quindi possa gestire situazioni di più etichette associate ad ogni osservazione.

Nota bene:

Con l'obiettivo di confrontare il risultato del clustering con le etichette disponibili, trasformare l'unica label ottenuta con il clustering, in formato **dummy**.

Esempio: il numero di cluster è 10 è il cluster predetto è etichettato con il numero **7** allora in formato dummy sarà **0 0 0 0 0 0 1 0 0 0**.

Un criterio utili allo scopo è l'indice di **Overlapped Cluster (OC)**¹ definito come il rapporto tra la probabilità di trovare due elementi raggruppati in entrambe le soluzioni e la massima probabilità di trovarli in una delle soluzioni date, essendo una probabilità assume valori da 0 ad 1, un valore vicino ad 1 definisce una forte equivalenza tra le due soluzioni, poiché qualsiasi coppia di oggetti può essere trovata in esse.

Dettagli sul criterio OC:

¹ CAMPO, David Nazareno; STEGMAYER, Georgina; MILONE, Diego H. A new index for clustering validation with overlapped clusters. *Expert Systems with Applications*, 2016, 64: 549-556.
<https://www.sciencedirect.com/science/article/abs/pii/S0957417416304158>

$$\mathcal{OC} = \frac{\tilde{t}}{\max(\tilde{p}, \tilde{p}')}.$$

Dove La probabilità di trovare due data points in entrambe le soluzioni viene stimata come :

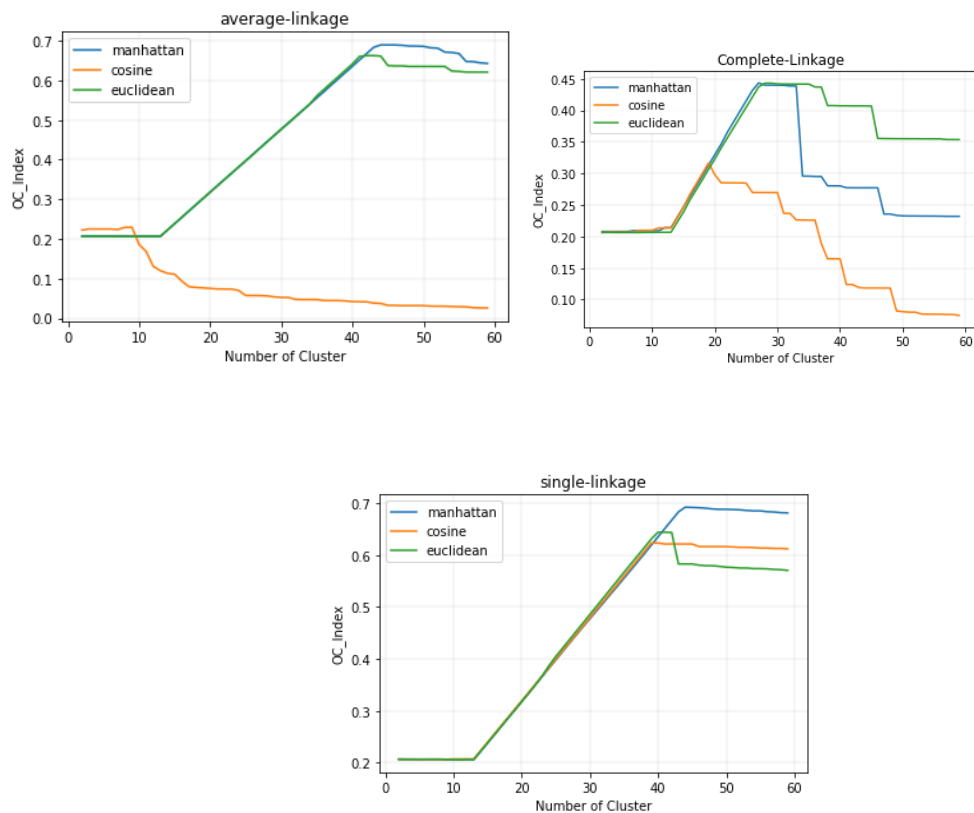
$$\tilde{t} = \frac{\sum_{i=1}^k \sum_{j=1}^{k'} \binom{|c_i \cap c'_j|}{2}}{\binom{N}{2} \frac{\max(n, n')}{N} \min(k, k')},$$

La probabilità di trovare una coppia di elementi in qualsiasi cluster per tutti i cluster viene stimata come

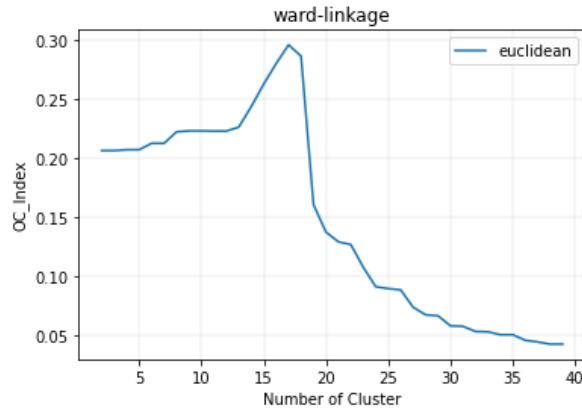
$$\tilde{p}' = \frac{\sum_{j=1}^{k'} \binom{|c'_j|}{2}}{k' \binom{N}{2}}.$$

OC INDEX

Utilizzando come criterio esterno OC avremo i seguenti risultati.



MYOCARDIAL INFARCTION-COMPLICATIONS



Il grafici mostrano le performance valutate con L'indice **OC** degli algoritmi di clustering agglomerativi Single-Linkage, Average-Linkage e Complete-Linkage al variare del numero K di cluster e della funzione di distanza utilizzata.

Nel caso dell'Average-Linkage la distanza di Mahnathan permette di ottenere performance migliori delle altre distanza che infatti da un certo numero di cluster tendono a decrescere verso lo o mentre con la Mahnathan si arriva al "gomito" dopo i 40 cluster per poi decrescere.

L'algoritmo Ward attiva al gomito ad un numero di clsuter pari a 17.

BUILD MODEL

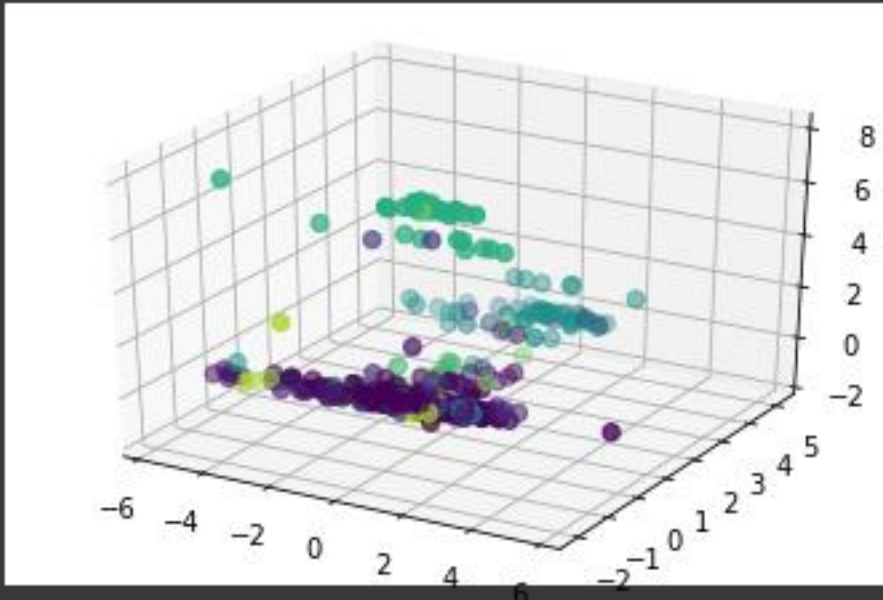
```
MODEL OUTPUT
=====
Ward, Affinity : euclidean, n_clusters : 17

OC Index : 0.2956842267375695
C_0 size : 1085 labels : {'A_V_BLOK': 26, 'DRESSLER': 46,
'FIBR_JELUD': 33, 'FIBR_PREDS': 68, 'JELUD_TAH': 18, 'LET_IS': 82,
'NO_COMP': 431, 'OTEK_LANC': 49, 'PREDS_TAH': 7, 'P_IM_STEN': 74,
'RAZRIV': 23, 'REC_IM': 65, 'ZSN': 163}
C_1 size : 440 labels : {'A_V_BLOK': 8, 'DRESSLER': 13, 'FIBR_JELUD':
7, 'FIBR_PREDS': 29, 'JELUD_TAH': 8, 'LET_IS': 75, 'NO_COMP': 95,
'OTEK_LANC': 46, 'PREDS_TAH': 4, 'P_IM_STEN': 28, 'RAZRIV': 12,
'REC_IM': 37, 'ZSN': 78}
C_2 size : 119 labels : {'A_V_BLOK': 4, 'DRESSLER': 1, 'FIBR_JELUD':
6, 'FIBR_PREDS': 11, 'JELUD_TAH': 2, 'LET_IS': 19, 'NO_COMP': 12,
'OTEK_LANC': 12, 'PREDS_TAH': 2, 'P_IM_STEN': 9, 'RAZRIV': 4,
'REC_IM': 12, 'ZSN': 25}
C_3 size : 44 labels : {'A_V_BLOK': 1, 'DRESSLER': 1, 'FIBR_JELUD': 1,
'FIBR_PREDS': 5, 'JELUD_TAH': 2, 'LET_IS': 10, 'NO_COMP': 7,
'OTEK_LANC': 5, 'P_IM_STEN': 1, 'RAZRIV': 1, 'REC_IM': 2, 'ZSN': 8}
C_4 size : 9 labels : {'FIBR_PREDS': 2, 'LET_IS': 1, 'NO_COMP': 2,
'OTEK_LANC': 1, 'PREDS_TAH': 3}
C_5 size : 14 labels : {'FIBR_JELUD': 1, 'FIBR_PREDS': 1, 'LET_IS': 1,
'NO_COMP': 3, 'OTEK_LANC': 1, 'P_IM_STEN': 2, 'REC_IM': 1, 'ZSN': 4}
C_6 size : 112 labels : {'A_V_BLOK': 2, 'DRESSLER': 4, 'FIBR_JELUD':
5, 'FIBR_PREDS': 6, 'JELUD_TAH': 3, 'LET_IS': 9, 'NO_COMP': 28,
'OTEK_LANC': 9, 'PREDS_TAH': 1, 'P_IM_STEN': 9, 'RAZRIV': 2, 'REC_IM':
10, 'ZSN': 24}
```

```

C_7 size : 58 labels : {'A_V_BLOK': 2, 'DRESSLER': 1, 'FIBR_JELUD': 3,
'FIBR_PREDS': 8, 'JELUD_TAH': 2, 'LET_IS': 5, 'NO_COMP': 20,
'OTEK_LANC': 1, 'P_IM_STEN': 2, 'RAZRIV': 2, 'REC_IM': 4, 'ZSN': 8}
C_8 size : 80 labels : {'A_V_BLOK': 2, 'DRESSLER': 2, 'FIBR_JELUD': 3,
'FIBR_PREDS': 17, 'JELUD_TAH': 1, 'LET_IS': 18, 'NO_COMP': 6,
'OTEK_LANC': 6, 'P_IM_STEN': 1, 'RAZRIV': 3, 'REC_IM': 3, 'ZSN': 18}
C_9 size : 4 labels : {'LET_IS': 1, 'NO_COMP': 2, 'P_IM_STEN': 1}
C_10 size : 43 labels : {'FIBR_JELUD': 2, 'FIBR_PREDS': 1, 'LET_IS':
12, 'NO_COMP': 4, 'OTEK_LANC': 5, 'P_IM_STEN': 1, 'RAZRIV': 2, 'ZSN':
16}
C_11 size : 7 labels : {'FIBR_PREDS': 1, 'NO_COMP': 3, 'OTEK_LANC': 1,
'ZSN': 2}
C_12 size : 11 labels : {'FIBR_PREDS': 1, 'JELUD_TAH': 1, 'NO_COMP':
1, 'OTEK_LANC': 2, 'REC_IM': 3, 'ZSN': 3}
C_13 size : 36 labels : {'A_V_BLOK': 1, 'DRESSLER': 2, 'FIBR_JELUD':
1, 'FIBR_PREDS': 3, 'LET_IS': 5, 'NO_COMP': 12, 'OTEK_LANC': 1,
'RAZRIV': 1, 'REC_IM': 1, 'ZSN': 9}
C_14 size : 35 labels : {'A_V_BLOK': 6, 'DRESSLER': 2, 'FIBR_JELUD':
3, 'FIBR_PREDS': 1, 'JELUD_TAH': 3, 'LET_IS': 9, 'NO_COMP': 2,
'OTEK_LANC': 2, 'P_IM_STEN': 1, 'RAZRIV': 3, 'REC_IM': 1, 'ZSN': 2}
C_15 size : 8 labels : {'FIBR_PREDS': 1, 'LET_IS': 1, 'OTEK_LANC': 2,
'REC_IM': 1, 'ZSN': 3}
C_16 size : 1 labels : {'ZSN': 1}

```



MODEL OUTPUT

```
=====
```

```
Complete-linkage, Affinity : euclidean, n_clusters : 28
```

```
OC_INDEX : 0.44304337764647056
```

```

C_0 size : 11 labels : {'FIBR_PREDS': 1, 'NO_COMP': 2, 'OTEK_LANC': 1,
'P_IM_STEN': 2, 'REC_IM': 1, 'ZSN': 4}
C_1 size : 119 labels : {'A_V_BLOK': 7, 'DRESSLER': 2, 'FIBR_JELUD':
6, 'FIBR_PREDS': 20, 'JELUD_TAH': 3, 'LET_IS': 15, 'NO_COMP': 33,
'OTEK_LANC': 3, 'PREDS_TAH': 1, 'P_IM_STEN': 3, 'RAZRIV': 4, 'REC_IM':
7, 'ZSN': 15}
C_2 size : 89 labels : {'A_V_BLOK': 2, 'DRESSLER': 2, 'FIBR_JELUD': 4,
'FIBR_PREDS': 17, 'JELUD_TAH': 2, 'LET_IS': 21, 'NO_COMP': 7,
'OTEK_LANC': 6, 'P_IM_STEN': 2, 'RAZRIV': 4, 'REC_IM': 3, 'ZSN': 19}
C_3 size : 43 labels : {'FIBR_JELUD': 2, 'FIBR_PREDS': 1, 'LET_IS':
12, 'NO_COMP': 4, 'OTEK_LANC': 5, 'P_IM_STEN': 1, 'RAZRIV': 2, 'ZSN':
16}

```

MYOCARDIAL INFARCTION-COMPLICATIONS

C_4 size : 52 labels : {'A_V_BLOK': 3, 'DRESSLER': 1, 'FIBR_JELUD': 1, 'FIBR_PREDS': 3, 'LET_IS': 10, 'NO_COMP': 5, 'OTEK_LANC': 8, 'P_IM_STEN': 3, 'RAZRIV': 2, 'REC_IM': 8, 'ZSN': 8}

C_5 size : 4 labels : {'LET_IS': 1, 'NO_COMP': 2, 'P_IM_STEN': 1}

C_6 size : 23 labels : {'FIBR_JELUD': 3, 'FIBR_PREDS': 4, 'JELUD_TAH': 2, 'LET_IS': 2, 'OTEK_LANC': 5, 'P_IM_STEN': 1, 'REC_IM': 2, 'ZSN': 4}

C_7 size : 39 labels : {'A_V_BLOK': 1, 'DRESSLER': 1, 'FIBR_PREDS': 5, 'JELUD_TAH': 2, 'LET_IS': 8, 'NO_COMP': 8, 'OTEK_LANC': 4, 'P_IM_STEN': 1, 'RAZRIV': 1, 'REC_IM': 1, 'ZSN': 7}

C_8 size : 21 labels : {'FIBR_JELUD': 1, 'FIBR_PREDS': 1, 'LET_IS': 3, 'NO_COMP': 7, 'OTEK_LANC': 2, 'P_IM_STEN': 2, 'REC_IM': 1, 'ZSN': 4}

C_9 size : 11 labels : {'FIBR_JELUD': 1, 'JELUD_TAH': 2, 'NO_COMP': 2, 'OTEK_LANC': 2, 'REC_IM': 1, 'ZSN': 3}

C_10 size : 8 labels : {'FIBR_PREDS': 1, 'LET_IS': 1, 'OTEK_LANC': 2, 'REC_IM': 1, 'ZSN': 3}

C_11 size : 1 labels : {'LET_IS': 1}

C_12 size : 68 labels : {'DRESSLER': 2, 'FIBR_JELUD': 3, 'FIBR_PREDS': 2, 'JELUD_TAH': 1, 'LET_IS': 6, 'NO_COMP': 10, 'OTEK_LANC': 9, 'PREDS_TAH': 1, 'P_IM_STEN': 7, 'REC_IM': 11, 'ZSN': 16}

C_13 size : 1494 labels : {'A_V_BLOK': 32, 'DRESSLER': 60, 'FIBR_JELUD': 36, 'FIBR_PREDS': 89, 'JELUD_TAH': 24, 'LET_IS': 147, 'NO_COMP': 527, 'OTEK_LANC': 88, 'PREDS_TAH': 11, 'P_IM_STEN': 105, 'RAZRIV': 35, 'REC_IM': 96, 'ZSN': 244}

C_14 size : 2 labels : {'FIBR_JELUD': 1, 'NO_COMP': 1}

C_15 size : 7 labels : {'FIBR_PREDS': 1, 'NO_COMP': 3, 'OTEK_LANC': 1, 'ZSN': 2}

C_16 size : 34 labels : {'A_V_BLOK': 6, 'DRESSLER': 2, 'FIBR_JELUD': 3, 'FIBR_PREDS': 1, 'JELUD_TAH': 3, 'LET_IS': 9, 'NO_COMP': 1, 'OTEK_LANC': 2, 'P_IM_STEN': 1, 'RAZRIV': 3, 'REC_IM': 1, 'ZSN': 2}

C_17 size : 4 labels : {'FIBR_JELUD': 1, 'LET_IS': 1, 'REC_IM': 1, 'ZSN': 1}

C_18 size : 4 labels : {'FIBR_PREDS': 1, 'NO_COMP': 1, 'OTEK_LANC': 1, 'PREDS_TAH': 1}

C_19 size : 1 labels : {'ZSN': 1}

C_20 size : 2 labels : {'LET_IS': 1, 'OTEK_LANC': 1}

C_21 size : 2 labels : {'ZSN': 2}

C_22 size : 4 labels : {'FIBR_JELUD': 1, 'FIBR_PREDS': 1, 'LET_IS': 1, 'PREDS_TAH': 1}

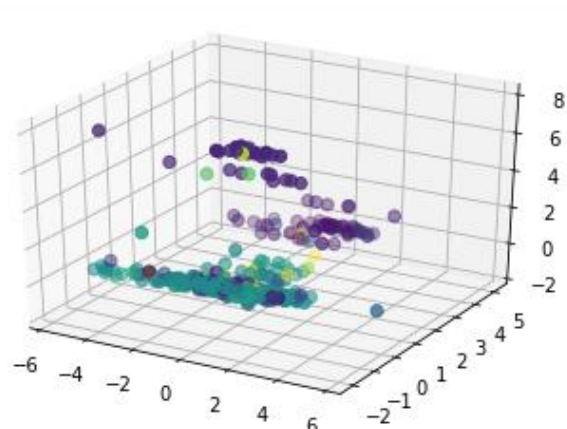
C_23 size : 3 labels : {'LET_IS': 1, 'RAZRIV': 1, 'REC_IM': 1}

C_24 size : 1 labels : {'OTEK_LANC': 1}

C_25 size : 5 labels : {'FIBR_PREDS': 1, 'LET_IS': 1, 'NO_COMP': 1, 'PREDS_TAH': 2}

C_26 size : 44 labels : {'A_V_BLOK': 1, 'DRESSLER': 2, 'FIBR_JELUD': 2, 'FIBR_PREDS': 5, 'LET_IS': 7, 'NO_COMP': 13, 'OTEK_LANC': 1, 'RAZRIV': 1, 'REC_IM': 2, 'ZSN': 10}

C_27 size : 10 labels : {'FIBR_PREDS': 1, 'JELUD_TAH': 1, 'NO_COMP': 1, 'OTEK_LANC': 1, 'REC_IM': 3, 'ZSN': 3}



MODEL OUTPUT

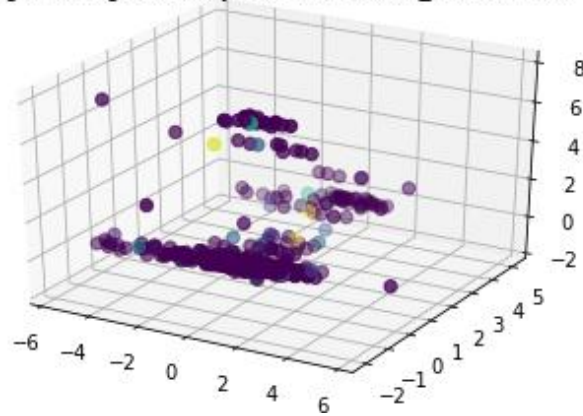
```
=====
singleLinkage, Affinity : manhattan, n_clusters : 42
```

```
OC_INDEX : 0.667212011912562
```

```
C_0 size : 2014 labels : {'A_V_BLOK': 51, 'DRESSLER': 72,
'FIBR_JELUD': 60, 'FIBR_PREDS': 146, 'JELUD_TAH': 38, 'LET_IS': 232,
'NO_COMP': 616, 'OTEK_LANC': 134, 'PREDS_TAH': 13, 'P_IM_STEN': 126,
'RAZRIV': 51, 'REC_IM': 132, 'ZSN': 343}
C_1 size : 3 labels : {'LET_IS': 2, 'ZSN': 1}
C_2 size : 2 labels : {'NO_COMP': 1, 'OTEK_LANC': 1}
C_3 size : 1 labels : {'LET_IS': 1}
C_4 size : 5 labels : {'NO_COMP': 3, 'OTEK_LANC': 1, 'ZSN': 1}
C_5 size : 5 labels : {'FIBR_PREDS': 1, 'LET_IS': 1, 'NO_COMP': 1,
'PREDS_TAH': 2}
C_6 size : 7 labels : {'NO_COMP': 1, 'OTEK_LANC': 1, 'P_IM_STEN': 1,
'REC_IM': 1, 'ZSN': 3}
C_7 size : 3 labels : {'LET_IS': 1, 'OTEK_LANC': 1, 'ZSN': 1}
C_8 size : 1 labels : {'LET_IS': 1}
C_9 size : 5 labels : {'JELUD_TAH': 1, 'NO_COMP': 1, 'OTEK_LANC': 1,
'REC_IM': 1, 'ZSN': 1}
C_10 size : 3 labels : {'FIBR_PREDS': 1, 'REC_IM': 1, 'ZSN': 1}
C_11 size : 1 labels : {'OTEK_LANC': 1}
C_12 size : 4 labels : {'FIBR_JELUD': 1, 'FIBR_PREDS': 1, 'LET_IS': 1,
'PREDS_TAH': 1}
C_13 size : 2 labels : {'FIBR_PREDS': 1, 'ZSN': 1}
C_14 size : 3 labels : {'LET_IS': 1, 'RAZRIV': 1, 'REC_IM': 1}
C_15 size : 1 labels : {'LET_IS': 1}
C_16 size : 4 labels : {'FIBR_PREDS': 1, 'LET_IS': 1, 'RAZRIV': 1,
'ZSN': 1}
C_17 size : 2 labels : {'REC_IM': 1, 'ZSN': 1}
C_18 size : 5 labels : {'FIBR_PREDS': 1, 'LET_IS': 1, 'OTEK_LANC': 1,
'REC_IM': 1, 'ZSN': 1}
C_19 size : 1 labels : {'LET_IS': 1}
C_20 size : 2 labels : {'LET_IS': 1, 'REC_IM': 1}
C_21 size : 1 labels : {'ZSN': 1}
C_22 size : 1 labels : {'OTEK_LANC': 1}
C_23 size : 1 labels : {'ZSN': 1}
C_24 size : 1 labels : {'P_IM_STEN': 1}
C_25 size : 1 labels : {'ZSN': 1}
C_26 size : 1 labels : {'P_IM_STEN': 1}
C_27 size : 1 labels : {'NO_COMP': 1}
C_28 size : 2 labels : {'FIBR_PREDS': 1, 'ZSN': 1}
C_29 size : 1 labels : {'NO_COMP': 1}
C_30 size : 1 labels : {'NO_COMP': 1}
C_31 size : 2 labels : {'OTEK_LANC': 1, 'ZSN': 1}
C_32 size : 1 labels : {'ZSN': 1}
C_33 size : 1 labels : {'FIBR_JELUD': 1}
C_34 size : 1 labels : {'NO_COMP': 1}
C_35 size : 3 labels : {'FIBR_JELUD': 1, 'FIBR_PREDS': 1, 'ZSN': 1}
C_36 size : 2 labels : {'FIBR_PREDS': 1, 'PREDS_TAH': 1}
C_37 size : 2 labels : {'A_V_BLOK': 1, 'ZSN': 1}
C_38 size : 1 labels : {'NO_COMP': 1}
C_39 size : 1 labels : {'LET_IS': 1}
C_40 size : 4 labels : {'FIBR_JELUD': 1, 'LET_IS': 1, 'REC_IM': 1,
'ZSN': 1}
C_41 size : 3 labels : {'FIBR_JELUD': 1, 'JELUD_TAH': 1, 'LET_IS': 1}
```

MYOCARDIAL INFARCTION-COMPLICATIONS

single-linkage, Affinity : manhattan, n_clusters : 42



MODEL OUTPUT

=====

averageLinkage, Affinity : manhattan, n_clusters : 42

OC_INDEX : 0.6675760355265243

C_0 size : 7 labels : {'FIBR_PREDS': 1, 'LET_IS': 1, 'NO_COMP': 2, 'OTEK_LANC': 1, 'PREDS_TAH': 2}

C_1 size : 2008 labels : {'A_V_BLOK': 51, 'DRESSLER': 72, 'FIBR_JELUD': 60, 'FIBR_PREDS': 146, 'JELUD_TAH': 38, 'LET_IS': 229, 'NO_COMP': 616, 'OTEK_LANC': 133, 'PREDS_TAH': 13, 'P_IM_STEN': 126, 'RAZRIV': 51, 'REC_IM': 132, 'ZSN': 341}

C_2 size : 2 labels : {'NO_COMP': 1, 'P_IM_STEN': 1}

C_3 size : 5 labels : {'LET_IS': 3, 'ZSN': 2}

C_4 size : 1 labels : {'LET_IS': 1}

C_5 size : 5 labels : {'NO_COMP': 3, 'OTEK_LANC': 1, 'ZSN': 1}

C_6 size : 7 labels : {'NO_COMP': 1, 'OTEK_LANC': 1, 'P_IM_STEN': 1, 'REC_IM': 1, 'ZSN': 3}

C_7 size : 1 labels : {'LET_IS': 1}

C_8 size : 5 labels : {'JELUD_TAH': 1, 'NO_COMP': 1, 'OTEK_LANC': 1, 'REC_IM': 1, 'ZSN': 1}

C_9 size : 2 labels : {'LET_IS': 1, 'REC_IM': 1}

C_10 size : 1 labels : {'OTEK_LANC': 1}

C_11 size : 3 labels : {'LET_IS': 1, 'OTEK_LANC': 1, 'ZSN': 1}

C_12 size : 1 labels : {'LET_IS': 1}

C_13 size : 4 labels : {'FIBR_JELUD': 1, 'FIBR_PREDS': 1, 'LET_IS': 1, 'PREDS_TAH': 1}

C_14 size : 1 labels : {'LET_IS': 1}

C_15 size : 2 labels : {'FIBR_PREDS': 1, 'ZSN': 1}

C_16 size : 3 labels : {'LET_IS': 1, 'RAZRIV': 1, 'REC_IM': 1}

C_17 size : 4 labels : {'FIBR_JELUD': 1, 'LET_IS': 1, 'REC_IM': 1, 'ZSN': 1}

C_18 size : 4 labels : {'FIBR_PREDS': 1, 'LET_IS': 1, 'RAZRIV': 1, 'ZSN': 1}

C_19 size : 3 labels : {'FIBR_JELUD': 1, 'JELUD_TAH': 1, 'LET_IS': 1}

C_20 size : 1 labels : {'LET_IS': 1}

C_21 size : 1 labels : {'NO_COMP': 1}

C_22 size : 2 labels : {'A_V_BLOK': 1, 'ZSN': 1}

C_23 size : 2 labels : {'FIBR_PREDS': 1, 'ZSN': 1}

C_24 size : 5 labels : {'FIBR_PREDS': 1, 'LET_IS': 1, 'OTEK_LANC': 1, 'REC_IM': 1, 'ZSN': 1}

C_25 size : 1 labels : {'LET_IS': 1}

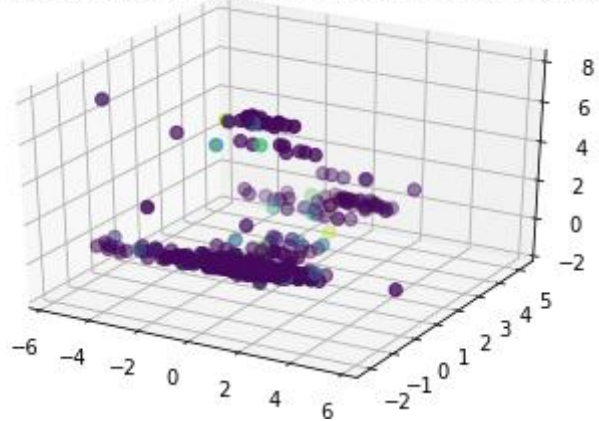
C_26 size : 1 labels : {'FIBR_JELUD': 1}

```

C_27 size : 3 labels : {'FIBR_JELUD': 1, 'FIBR_PREDS': 1, 'ZSN': 1}
C_28 size : 1 labels : {'OTEK_LANC': 1}
C_29 size : 2 labels : {'OTEK_LANC': 1, 'ZSN': 1}
C_30 size : 1 labels : {'P_IM_STEN': 1}
C_31 size : 1 labels : {'ZSN': 1}
C_32 size : 1 labels : {'NO_COMP': 1}
C_33 size : 2 labels : {'REC_IM': 1, 'ZSN': 1}
C_34 size : 1 labels : {'ZSN': 1}
C_35 size : 1 labels : {'ZSN': 1}
C_36 size : 3 labels : {'LET_IS': 1, 'OTEK_LANC': 1, 'ZSN': 1}
C_37 size : 3 labels : {'FIBR_PREDS': 1, 'REC_IM': 1, 'ZSN': 1}
C_38 size : 1 labels : {'NO_COMP': 1}
C_39 size : 1 labels : {'ZSN': 1}
C_40 size : 2 labels : {'FIBR_PREDS': 1, 'PREDS_TAH': 1}
C_41 size : 1 labels : {'NO_COMP': 1}

```

average-linkage, Affinity : manhattan, n_clusters : 42



ESTRAZIONE DI REGOLE DI ASSOCIAZIONE

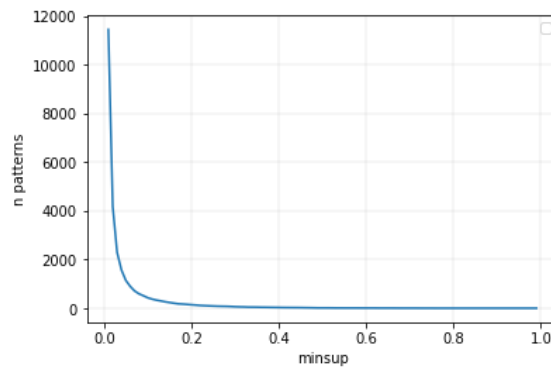
Utilizzando le variabili binarie è stato effettuato un task di Frequent Pattern Mining, applicando l'algoritmo **Apriori** con l'obiettivo a trovare pattern frequenti e regole di associazioni.

Il dataset ristretto alle soli variabili binari viene quindi considerato come un insieme di transazioni.

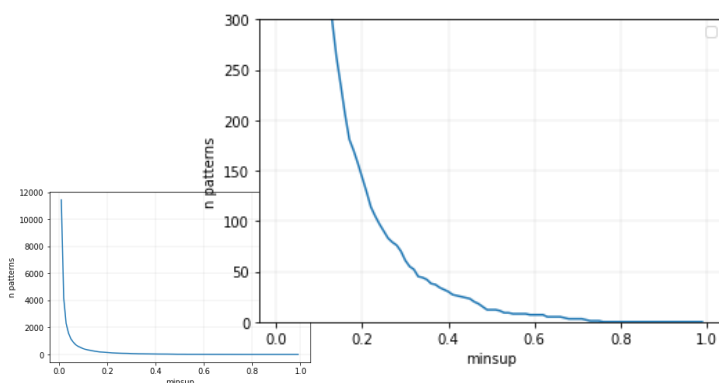
MYOCARDIAL INFARCTION-COMPLICATIONS

SEX	SIM_GIPERT	nr_11	nr_01	nr_02	nr_03	nr_04	nr_07	nr_08	np_01	np_04	np_05	np_07	np_08	np_09	np_10	endocr_01	endocr_02	endocr_03	zab_leg_01	zab_leg_02
1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0

Eseguendo l'algoritmo Apriori per diversi valori di **minsup** si ottiene il seguente grafico:



Il grafico mostra il numero di pattern scoperti nel dataset al variare della soglia di supporto minimo.



Osservando il grafico ingrandito si nota che ad una percentuale di soglia di supporto maggiore del 20 % i pattern tendono a diminuire drasticamente sino ad arrivare a non scoprire nessun pattern.

Si optato quindi per una percentuale di soglia di **minsup del 30%** e una soglia di **minconf del 70%**

I pattern frequenti scoperti sono i seguenti:

	frequent pattern	support
0	('ASPSn',)	1182
1	('LIDSn',)	454
2	('GEPARSn',)	1123
3	('ritmecgp_07',)	350
4	('SEX',)	986
5	('ANTCAS_n',)	1060
6	('ritmecgp_01',)	1052
7	('TRENTSn',)	315
8	('ZSN',)	364
9	('LET_IS',)	248
10	('ASPSn', 'GEPARSn')	923
11	('ASPSn', 'LIDSn')	355
12	('ASPSn', 'SEX')	732
13	('ASPSn', 'ritmecgp_07')	268
14	('GEPARSn', 'LIDSn')	342
15	('GEPARSn', 'SEX')	716
16	('GEPARSn', 'ritmecgp_07')	246
17	('LIDSn', 'SEX')	321
18	('ANTCASn', 'ASPS_n')	827
19	('ANTCASn', 'GEPARS_n')	758
20	('ANTCASn', 'LIDS_n')	301
21	('ANTCAS_n', 'SEX')	655
22	('ANTCASn', 'ritmecgp01')	719
23	('ASPSn', 'ritmecgp_01')	808
24	('GEPARSn', 'ritmecgp_01')	767
25	('LIDSn', 'ritmecgp_01')	312
26	('SEX', 'ritmecgp_01')	700
27	('ANTCAS_n', 'ZSN')	258
28	('ASPSn', 'ZSN')	276
29	('GEPARSn', 'ZSN')	249
30	('ASPSn', 'GEPARSn', 'LIDSn')	291
31	('ASPSn', 'GEPARSn', 'SEX')	594
32	('ASPSn', 'LIDSn', 'SEX')	246
33	('GEPARSn', 'LIDSn', 'SEX')	245
34	('ANTCASn', 'ASPSn', 'GEPARS_n')	627
35	('ANTCASn', 'ASPSn', 'LIDS_n')	240
36	('ANTCASn', 'ASPS_n', 'SEX')	497
37	('ANTCASn', 'ASPSn', 'ritmecgp01')	565
38	('ANTCASn', 'GEPARS_n', 'SEX')	472
39	('ANTCASn', 'GEPARSn', 'ritmecgp01')	517

MYOCARDIAL INFARCTION-COMPLICATIONS

40	('ANTCAsn', 'SEX', 'ritmecgp01')	472
41	('ASPSn', 'GEPARSn', 'ritmecgp_01')	640
42	('ASPSn', 'LIDSn', 'ritmecgp_01')	243
43	('ASPSn', 'SEX', 'ritmecgp_01')	535
44	('GEPARSn', 'LIDSn', 'ritmecgp_01')	244
45	('GEPARSn', 'SEX', 'ritmecgp_01')	517
46	('ANTCAsn', 'ASPSn', 'GEPARSn', 'SEX')	392
47	('ANTCAsn', 'ASPSn', 'GEPARSn', 'ritmecgp01')	433
48	('ANTCAsn', 'ASPSn', 'SEX', 'ritmecgp01')	365
49	('ANTCAsn', 'GEPARSn', 'SEX', 'ritmecgp01')	344
50	('ASPSn', 'GEPARSn', 'SEX', 'ritmecgp_01')	438
51	('ANTCAsn', 'ASPSn', 'GEPARSn', 'SEX', 'ritmecgp01')	292

Le regole di associazione che superano una soglia di minconf dell 70% sono:

```
[{GEPAR_S_n} -> {ASP_S_n},  
{ASP_S_n} -> {GEPAR_S_n},  
{LID_S_n} -> {ASP_S_n},  
{SEX} -> {ASP_S_n},  
{ritm_ecg_p_07} -> {ASP_S_n},  
{LID_S_n} -> {GEPAR_S_n},  
{SEX} -> {GEPAR_S_n},  
{ritm_ecg_p_07} -> {GEPAR_S_n},  
{LID_S_n} -> {SEX},  
{ANT_CA_S_n} -> {ASP_S_n},  
{ANT_CA_S_n} -> {GEPAR_S_n},
```

{ritm_ecg_p_01} -> {ASP_S_n},
{ritm_ecg_p_01} -> {GEPAR_S_n},
{SEX} -> {ritm_ecg_p_01},
{ZSN} -> {ANT_CA_S_n} ,
{ZSN} ->{ASP_S_n} ,
{GEPAR_S_n, LID_S_n} -> {ASP_S_n},
{ASP_S_n, LID_S_n} -> {GEPAR_S_n},
{GEPAR_S_n, SEX} -> {ASP_S_n},
{ASP_S_n, SEX} -> {GEPAR_S_n},
{LID_S_n, SEX} -> {ASP_S_n},
{LID_S_n, SEX} -> {GEPAR_S_n},
{GEPAR_S_n, LID_S_n} -> {SEX},
{ANT_CA_S_n, GEPAR_S_n} -> {ASP_S_n},
{ANT_CA_S_n, ASP_S_n} -> {GEPAR_S_n},
{ANT_CA_S_n, LID_S_n} -> {ASP_S_n},
{ANT_CA_S_n, SEX} -> {ASP_S_n},
{ANT_CA_S_n, ritm_ecg_p_01} -> {ASP_S_n},
{ANT_CA_S_n, SEX} -> {GEPAR_S_n},
{ANT_CA_S_n, ritm_ecg_p_01} -> {GEPAR_S_n},
{ANT_CA_S_n, SEX} -> {ritm_ecg_p_01},
{GEPAR_S_n, ritm_ecg_p_01} -> {ASP_S_n},
{ASP_S_n, ritm_ecg_p_01} -> {GEPAR_S_n},
{LID_S_n, ritm_ecg_p_01} -> {ASP_S_n},
{SEX, ritm_ecg_p_01} -> {ASP_S_n},
{ASP_S_n, SEX} -> {ritm_ecg_p_01},
{LID_S_n, ritm_ecg_p_01} -> {GEPAR_S_n},
{GEPAR_S_n, LID_S_n} -> {ritm_ecg_p_01},
{SEX, ritm_ecg_p_01} -> {GEPAR_S_n},
{GEPAR_S_n, SEX} -> {ritm_ecg_p_01},

MYOCARDIAL INFARCTION-COMPLICATIONS

```
{ANT_CA_S_n, GEPAR_S_n, SEX} -> {ASP_S_n},
{ANT_CA_S_n, ASP_S_n, SEX} -> {GEPAR_S_n},
{ANT_CA_S_n, GEPAR_S_n, ritm_ecg_p_01} -> {ASP_S_n},
{ANT_CA_S_n, ASP_S_n, ritm_ecg_p_01} -> {GEPAR_S_n},
{ANT_CA_S_n, SEX, ritm_ecg_p_01} -> {ASP_S_n},
{ANT_CA_S_n, ASP_S_n, SEX} -> {ritm_ecg_p_01},
{ANT_CA_S_n, SEX, ritm_ecg_p_01} -> {GEPAR_S_n},
{ANT_CA_S_n, GEPAR_S_n, SEX} -> {ritm_ecg_p_01},
{GEPAR_S_n, SEX, ritm_ecg_p_01} -> {ASP_S_n},
{ASP_S_n, SEX, ritm_ecg_p_01} -> {GEPAR_S_n},
{ASP_S_n, GEPAR_S_n, SEX} -> {ritm_ecg_p_01},
{ANT_CA_S_n, GEPAR_S_n, SEX, ritm_ecg_p_01} -> {ASP_S_n},
{ANT_CA_S_n, ASP_S_n, SEX, ritm_ecg_p_01} -> {GEPAR_S_n},
{ANT_CA_S_n, ASP_S_n, GEPAR_S_n, SEX} -> {ritm_ecg_p_01}]
```

EVALUATION

ANALISI DEI CLUSTER:

Utilizzando i **criteri interni** si giunge alla conclusione che per tutti gli algoritmi utilizzati il numero di **cluster naturali** è **2** con :

- Average Linkage con distanza di manhattan
- Complete Linkage con distanza euclidean
- Ward

E 350 con Average Linkage e distanza del coseno.

In generale si va a formare **un grande cluster** comprendente quasi tutti gli esempi e **un piccolo cluster** contenente un numero ridotto di esempi.

Con l'Average-Linkage con metrica coseno e Ward si ottiene un indice di silhouette inferiore rispetto agli altri però dal dendrogramma sembrerebbe che il raggruppamento in cluster sia più equilibrato.

Utilizzando come **criterio esterno OC** si ottiene una soluzione con un numero intermedio di cluster, le soluzioni trovate sono caratterizzate da **un grande cluster** e tanti **piccoli cluster di dimensioni ridotte**.

Da questi risultati si può concludere che si può individuare:

- un grande sottogruppo di pazienti considerati simili dove che alcuni presentano diversi tipi di complicazioni mentre altri non ne presentano nessuna
- e diversi sottogruppi più piccoli e più isolati di pazienti

Difficile quindi stabilire da questi risultati l'esistenza una qualche tipo di relazione con le complicazioni dei pazienti, in quanto le assegnazioni delle osservazioni ai cluster naturali sembri non corrispondere molto alle label del dataset.

Probabilmente **aumentando la dimensione dei dati** in modo da bilanciare il dataset sarà possibile ottenere risultati più interessanti con la cluster analysis

REGOLE DI ASSOCIAZIONE ESTRATTE:

Nell'estrazione di regole di associazioni sono state rilevate delle regole interessanti tra le regole estratte:

{ritm_ecg_p_07} -> {GEPAR_S_n}

Ai pazienti che hanno un Ritmo ECG al momento dell'ammissione in ospedale - seno con una frequenza cardiaca superiore a 90 (tachicardia) vengono successivamente somministrati degli anticoagulanti (eparina) in terapia intensiva.

{ritm_ecg_p_07} -> {ASP_S_n}

A i pazienti che hanno un Ritmo ECG al momento dell'ammissione in ospedale - seno con una frequenza cardiaca superiore a 90 (tachicardia) vengono somministrati Uso dell'acido acetilsalicilico in terapia intensiva

MYOCARDIAL INFARCTION-COMPLICATIONS

{ZSN} -> {ANT_CA_S_n} ,

Ai pazienti che hanno avuto complicazioni di Insufficienza cardiaca cronica è stato somministrato l'uso di calcio-antagonisti in terapia intensiva

{ZSN} ->{ASP_S_n}

Ai pazienti che hanno avuto complicazioni di Insufficienza cardiaca cronica è stato somministrato l'uso dell'acido acetilsalicilico in terapia intensiva