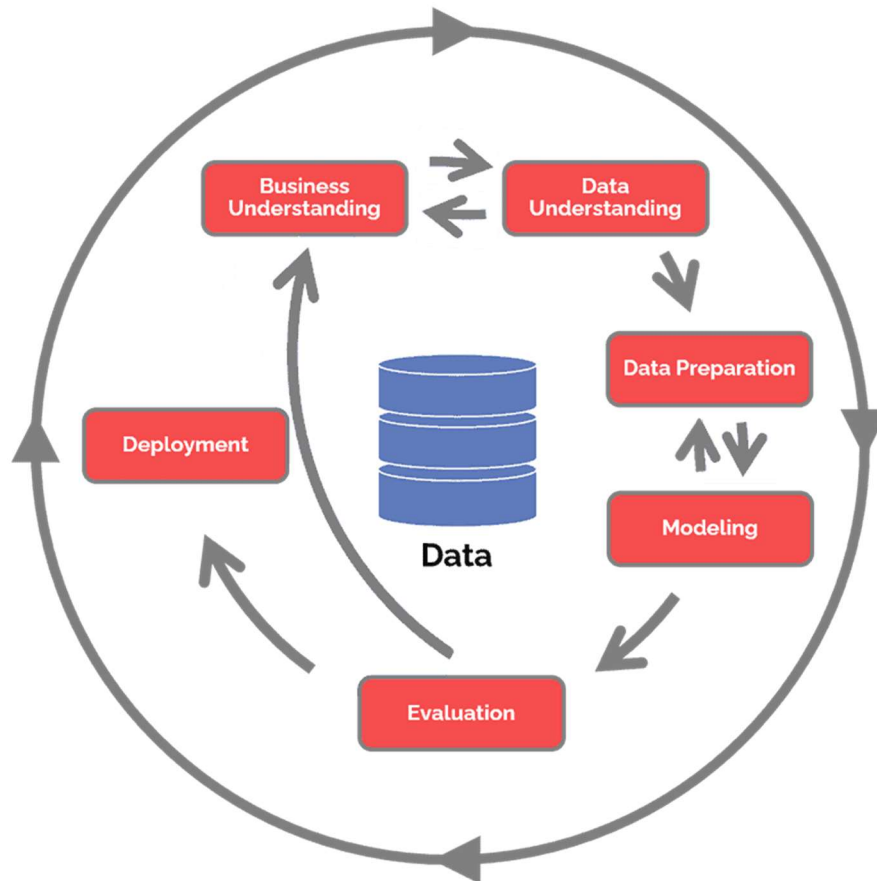


MYOCARDIAL INFARCTION- COMPLICATIONS ANALYSIS

ANALISI PER LA FENOTIPIZZAZIONE DELLA MALATTIA
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.

MYOCARDIAL INFARCTION-COMPLICATIONS

OBBIETTIVI DI BUSINESS

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

OBBIETTIVI DI DATA MINING

1. Analisi dei Cluster, individuare gruppi di pazienti con tra loro simili sulla base delle informazioni prese in considerazione con l'obiettivo di determinare una relazione con le complicazioni della malattia che i pazienti presentano.
1. Association Rules Extraction , estrazione di regole di associazione per estrarre relazioni nascose 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.

MYOCARDIAL INFARCTION-COMPLICATIONS

VARIABILI DI OUTPUT (TIPI DI COMPLICAZIONI)

BINARIE

MYOCARDIAL INFARCTION-COMPLICATIONS

Nome	Descrizione	Diagramma									
FIBR_PREDS	Atrial fibrillation	<table border="1"> <thead> <tr> <th>Category</th> <th>Value</th> <th>Percentage</th> </tr> </thead> <tbody> <tr> <td>Green</td> <td>0</td> <td>90.000%</td> </tr> <tr> <td>Blue</td> <td>1</td> <td>10.000%</td> </tr> </tbody> </table>	Category	Value	Percentage	Green	0	90.000%	Blue	1	10.000%
Category	Value	Percentage									
Green	0	90.000%									
Blue	1	10.000%									
PREDS_TAH	Supraventricular tachycardia	<table border="1"> <thead> <tr> <th>Category</th> <th>Value</th> <th>Percentage</th> </tr> </thead> <tbody> <tr> <td>Green</td> <td>0</td> <td>99.820%</td> </tr> <tr> <td>Blue</td> <td>1</td> <td>1.180%</td> </tr> </tbody> </table>	Category	Value	Percentage	Green	0	99.820%	Blue	1	1.180%
Category	Value	Percentage									
Green	0	99.820%									
Blue	1	1.180%									
JELUD_TAH	Ventricular tachycardia	<table border="1"> <thead> <tr> <th>Category</th> <th>Value</th> <th>Percentage</th> </tr> </thead> <tbody> <tr> <td>Green</td> <td>0</td> <td>97.530%</td> </tr> <tr> <td>Blue</td> <td>1</td> <td>2.470%</td> </tr> </tbody> </table>	Category	Value	Percentage	Green	0	97.530%	Blue	1	2.470%
Category	Value	Percentage									
Green	0	97.530%									
Blue	1	2.470%									
FIBR_JELUD	Ventricular fibrillation	<table border="1"> <thead> <tr> <th>Category</th> <th>Value</th> <th>Percentage</th> </tr> </thead> <tbody> <tr> <td>Green</td> <td>0</td> <td>95.820%</td> </tr> <tr> <td>Blue</td> <td>1</td> <td>4.180%</td> </tr> </tbody> </table>	Category	Value	Percentage	Green	0	95.820%	Blue	1	4.180%
Category	Value	Percentage									
Green	0	95.820%									
Blue	1	4.180%									
A_V_BLOK)	Third-degree AV block	<table border="1"> <thead> <tr> <th>Category</th> <th>Value</th> <th>Percentage</th> </tr> </thead> <tbody> <tr> <td>Green</td> <td>0</td> <td>96.650%</td> </tr> <tr> <td>Blue</td> <td>1</td> <td>3.350%</td> </tr> </tbody> </table>	Category	Value	Percentage	Green	0	96.650%	Blue	1	3.350%
Category	Value	Percentage									
Green	0	96.650%									
Blue	1	3.350%									

MYOCARDIAL INFARCTION-COMPLICATIONS

Nome	Descrizione	Diagramma
OTEK_LANC	Pulmonary edema	<p>90.650% 9.350% 1</p>
RAZRIV	Myocardial rupture	<p>96.820% 3.180% 1</p>
DRESSLER	Dressler syndrome	<p>95.590% 4.410% 1</p>
ZSN	Chronic heart failure	<p>76.820% 23.180% 1</p>
REC_IM	Relapse of the myocardial infarction	<p>90.650% 9.350% 1</p>

MYOCARDIAL INFARCTION-COMPLICATIONS

Variabili Qualitative (Categoriche Nominali)

Nome	Descrizione	Valori Assunti	Frazione
LET_IS	Esito mortale (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)

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

Cases	FIB R_ PR ED S	PR ED S_ TA H	JEL UD _T AH UD	FIB R_J EL UD	A_ V_ BL OK	OT EK _L AN	RA ZRI V	DR ES SL ER	ZS N	RE C_I M	P_I M_ ST EN	LE T_I S
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

MYOCARDIAL INFARCTION-COMPLICATIONS

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

FIB	PR	JEL	FIB	A_	OT	RA	DR	ZS	RE	P_I	LE
R_	ED	UD	R_J	V_	EK	ZR	ES	N	C_I	M_	T_I
PR	S_	_T	EL	BL	_L	IV	SL		M	ST	S
ED	TA	AH	UD	OK	AN		ER			EN	
S	H										

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
1				X					X	X			X
3				X			X						X
1				X		X			X	X			
1				X		X		X		X			

MYOCARDIAL INFARCTION-COMPLICATIONS

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							
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	FIB R_ PR ED S	PR ED S_ TA H	JEL UD _T AH	FIB R_ EL UD	A_ V_ BL OK	OT EK _L AN	RA ZR IV	DR ES SL ER	ZS N	RE C_ M	P_ I M_ ST EN	LE T_ S
30		X							X			
2		X							X			X
1		X							X		X	
4		X							X	X		
1		X							X	X		X
1		X						X			X	

MYOCARDIAL INFARCTION-COMPLICATIONS

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			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

	FIB R_ PR ED S	PR ED S_ TA H	JEL UD _T AH	FIB R_J EL UD	A_ V_ BL OK	OT EK _L AN	RA ZRI V	DR ES SL ER	ZS N	RE C_I M	P_I M_ ST EN	LE T_I S	Co m pli cat io
4						X			X	X		X	4
1						X		X	X	X			4
1						X	X			X		X	4
1						X	X		X			X	4
1					X	X			X	X			4
1				X					X	X		X	4
1				X		X			X	X			4
1				X		X		X		X			4

MYOCARDIAL INFARCTION-COMPLICATIONS

1			X		X		X	X				4
1			X	X		X					X	4
1			X				X		X	X		4
1			X			X			X		X	4
1			X			X		X				4
1			X	X				X	X			4
1			X	X			X		X			4
1	X							X	X		X	4
1	X						X	X			X	4
1	X						X	X		X		4
1	X					X			X		X	4
2	X					X		X			X	4
1	X					X	X				X	4
1	X				X			X			X	4
1	X			X		X					X	4
1	X			X	X						X	4
1	X			X					X		X	4
2	X			X				X	X			4
1	X		X				X				X	4
1	X		X		X						X	4
1	X		X	X		X						4
1	X	X					X	X				4
1	X	X		X							X	4
1					X		X	X	X		X	5
1			X	X	X			X			X	5
1	X				X			X	X		X	5
1	X		X		X			X	X			5
1	X		X	X					X		X	5
1	X		X	X			X	X				5
1	X	X			X			X			X	5

MYOCARDIAL INFARCTION-COMPLICATIONS

VARIABILI DI INPUT

QUANTITATIVE

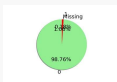




Name	Descrizione	Missing	Q1	Median	Q3	Min	Mean	Max	Std
AGE		8	54	63	70	26	61.857	92	11.2566
S_AD_KBRIG	Pressione sanguigna sistolica secondo il team di cardiologia d'emergenza	1076	120	140	160	0	136.907	260	34.9698
D_AD_KBRIG		1076	70	80	90	0	81.3942	190	19.7292
S_AD_ORIT		267	120	130	150	0	134.588	260	31.3374
D_AD_ORIT		267	80	80	90	0	82.7495	190	18.3147
NA_BLOOD		375	133	136	140	117	136.551	169	6.50966
ALT_BLOOD		284	0.23	0.38	0.61	0.03	0.481455	3	0.387124
AST_BLOOD		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		125	6.4	8	10.45	2	8.78291	27.9	3.39948
K_BLOOD			3.7	4.1	4.6	2.3	4.19	8.2	0.75
ROE		203	5	10	18	1	13.4449	140	11.2925


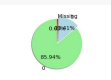



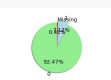

Categoriche Booleane

MYOCARDIAL INFARCTION-COMPLICATIONS





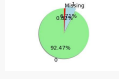


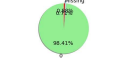

Name	Description	1	o	Missings	Pie
SEX	Genere del pazione o:Donna 1:Uomo	62.65%	37.35%	0.0%	
IBS_NASL	Ereditarietà su CHD. o:non presente 1 : presente	1.59%	2.65%	95.76%	
SIM_GIPERT	Presenza di ipertensione sintomatica	3.35%	96.18%	0.47%	
nr_11	Osservazione dell'aritmia nell'anamnesi	2.47%	96.29%	1.24%	
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%	




MYOCARDIAL INFARCTION-COMPLICATIONS

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%	
zab_leg_02	Bronchite cronica ostruttiva nell'anamnesi	7.12%	92.47%	0.41%	
zab_leg_03	Asma bronchiale nell'anamnesi	2.18%	97.41%	0.41%	

MYOCARDIAL INFARCTION-COMPLICATIONS

zab_leg_04		0.53%	99.06%	0.41%	
zab_leg_06	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	0	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%	





MYOCARDIAL INFARCTION-COMPLICATIONS

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


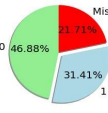






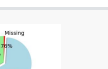

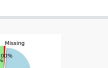
MYOCARDIAL INFARCTION-COMPLICATIONS

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 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 Celasum 750k IU	0.76%	98.65%	0.59%	
fibr_ter_02	Terapia fibrinolitica con Celasum 1m IU	0.94%	98.47%	0.59%	



MYOCARDIAL INFARCTION-COMPLICATIONS

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

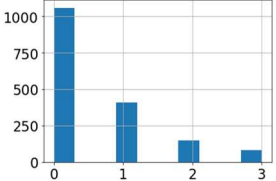
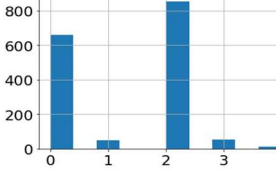
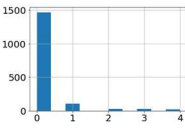
Name	Description	1	o	Missings	Pie
fibr_ter_o8	Terapia fibrinolitica con Streptodecase 1,5m IU	0.12%	99.29%	0.59%	
GIPO_K	Ipopotassiemia (< 4 mmol/L)	31.41%	46.88%	21.71%	
GIPER_NA	Aumento del sodio nel siero	1.76%	76.18%	22.06%	
NA_KB	Uso di farmaci oppioidi da parte del team di cardiologia d'emergenza	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 acetilsalicilico in terapia intensiva	73.65%	25.35%	1.0%	

MYOCARDIAL INFARCTION-COMPLICATIONS

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	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	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)	

MYOCARDIAL INFARCTION-COMPLICATIONS

ant_im	Presenza di un infarto miocardico anteriore (ventricolare sinistro) (cambiamenti ECG nelle derivazioni V1 - V4)	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)	
lat_im	Presenza di un infarto miocardico laterale (ventricolare sinistro) (cambiamenti ECG nelle derivazioni V5 - V6, I, AVL)	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)	

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)	
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)	
TIME_B_S	Tempo trascorso dall'inizio dell'attacco di CHD al ospedale	1: medio 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)	

MYOCARDIAL INFARCTION-COMPLICATIONS

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)	

MYOCARDIAL INFARCTION-COMPLICATIONS

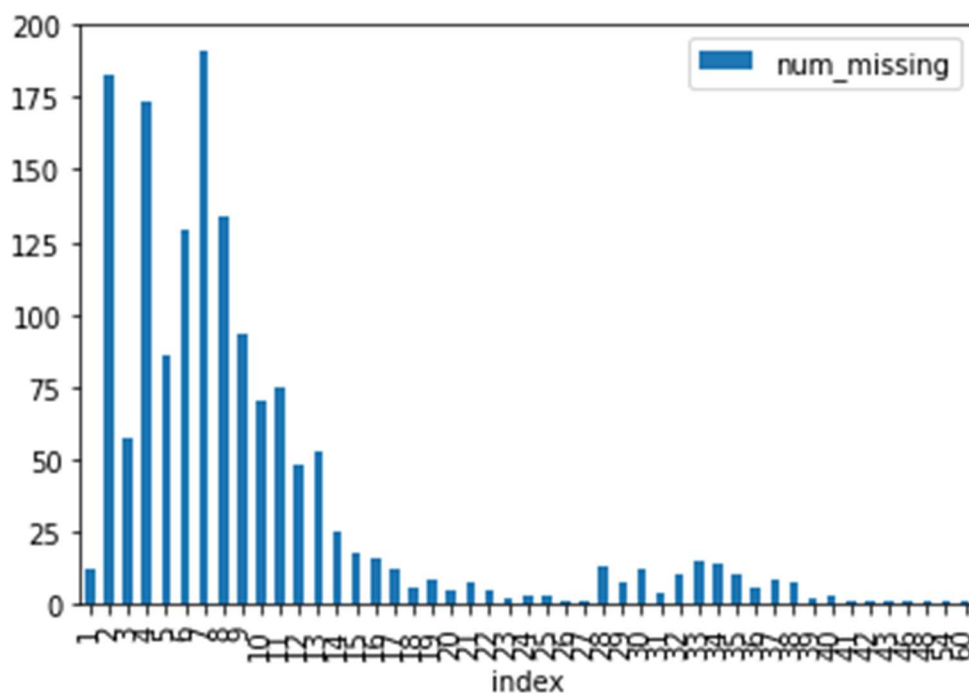
DATA PREPARATION

PROBLEMI RISCONTRATI

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

MISSING DATA HISTOGRAM

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



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

LISTA DI PERCENTUALE DI DATI MANCANTI

ID - 0%	JELUD_TAH - 0%	RAZRIV - 0%
SEX - 0%	FIBR_JELUD - 0%	DRESSLER - 0%
FIBR_PREDS - 0%	A_V_BLOK - 0%	ZSN - 0%
PREDS_TAH - 0%	OTEK_LANC - 0%	REC_IM - 0%

MYOCARDIAL INFARCTION-COMPLICATIONS

P_IM_STEN - 0%	K_SH_POST - 1%	n_p_ecg_p_04 - 7%
LET_IS - 0%	R_AB_1_n - 1%	n_p_ecg_p_05 - 7%
IM_PG_P - 0%	TIKL_S_n - 1%	n_p_ecg_p_06 - 7%
INF_ANAM - 0%	TRENT_S_n - 1%	n_p_ecg_p_07 - 7%
NA_R_1_n - 0%	np_01 - 1%	n_p_ecg_p_08 - 7%
zab_leg_01 - 0%	np_04 - 1%	n_p_ecg_p_09 - 7%
zab_leg_02 - 0%	np_05 - 1%	n_p_ecg_p_10 - 7%
zab_leg_03 - 0%	np_07 - 1%	n_p_ecg_p_11 - 7%
zab_leg_04 - 0%	np_08 - 1%	n_p_ecg_p_12 - 7%
zab_leg_06 - 0%	np_09 - 1%	n_r_ecg_p_01 - 7%
AGE - 1%	np_10 - 1%	n_r_ecg_p_02 - 7%
SIM_GIPERT - 1%	GEPAR_S_n - 1%	n_r_ecg_p_03 - 7%
NITR_S - 1%	ASP_S_n - 1%	n_r_ecg_p_04 - 7%
NOT_NA_1_n - 1%	nr_11 - 1%	n_r_ecg_p_05 - 7%
GB - 1%	nr_01 - 1%	n_r_ecg_p_06 - 7%
fibr_ter_01 - 1%	nr_02 - 1%	n_r_ecg_p_08 - 7%
fibr_ter_02 - 1%	nr_03 - 1%	n_r_ecg_p_09 - 7%
fibr_ter_03 - 1%	nr_04 - 1%	n_r_ecg_p_10 - 7%
fibr_ter_05 - 1%	nr_07 - 1%	R_AB_3_n - 7%
fibr_ter_06 - 1%	nr_08 - 1%	NA_R_3_n - 7%
fibr_ter_07 - 1%	IBS_POST - 2%	NOT_NA_3_n - 7%
fibr_ter_08 - 1%	ZSN_A - 3%	TIME_B_S - 7%
endocr_01 - 1%	FK_STENOK - 4%	L_BLOOD - 8%
endocr_02 - 1%	post_im - 4%	ritm_ecg_p_01 - 9%
endocr_03 - 1%	lat_im - 4%	ritm_ecg_p_02 - 9%
O_L_POST - 1%	ant_im - 5%	ritm_ecg_p_04 - 9%
SVT_POST - 1%	inf_im - 5%	ritm_ecg_p_06 - 9%
GT_POST - 1%	STENOK_AN - 6%	ritm_ecg_p_07 - 9%
FIB_G_POST - 1%	R_AB_2_n - 6%	ritm_ecg_p_08 - 9%
LID_S_n - 1%	NA_R_2_n - 6%	ROE - 12%
B_BLOK_S_n - 1%	NOT_NA_2_n - 6%	DLIT_AG - 15%
MP_TP_POST - 1%	n_p_ecg_p_01 - 7%	S_AD_ORIT - 17%
ANT_CA_S_n - 1%	n_p_ecg_p_03 - 7%	D_AD_ORIT - 17%

MYOCARDIAL INFARCTION-COMPLICATIONS

ALT_BLOOD - 17%	NOT_NA_KB - 41%
AST_BLOOD - 17%	S_AD_KBRIG - 63%
GIPO_K - 21%	D_AD_KBRIG - 63%
K_BLOOD - 22%	IBS_NASL - 96%
GIPER_NA - 22%	KFK_BLOOD - 100%
NA_BLOOD - 22%	
NA_KB - 39%	
LID_KB - 40%	

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)

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

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 feature che non mancano a nessuno dei due sono vicine, è stato settato un numero di vicini $K = 3$.

Frammento del dataset prima dell'applicazione del metodo:

MYOCARDIAL INFARCTION-COMPLICATIONS

GT_POST	...	K_BLOOD	GIPER_NA	NA_BLOOD	ALT_BLOOD	AST_BLOOD	L_BLOOD	ROE	TIME_B_S	R_AB_1_n	R_AB_2_n	R_AB_3_n	NA_KB	NOT_NA_KB	LID_KB	NITR_S	N
0.0	...	4.7	0.0	138.0	NaN	NaN	8.0	16.0	4.0	0.0	0.0	1.0	NaN	NaN	NaN	0.0	
0.0	...	3.5	0.0	132.0	0.38	0.18	7.8	3.0	2.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	
0.0	...	4.0	0.0	132.0	0.30	0.11	10.8	NaN	3.0	3.0	0.0	0.0	1.0	1.0	1.0	0.0	
0.0	...	3.9	0.0	146.0	0.75	0.37	NaN	NaN	2.0	0.0	0.0	1.0	NaN	NaN	NaN	0.0	
0.0	...	3.5	0.0	132.0	0.45	0.22	8.3	NaN	9.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
0.0	...	NaN	NaN	NaN	0.45	0.22	7.2	2.0	2.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	
0.0	...	NaN	NaN	NaN	0.30	0.11	11.1	5.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	
0.0	...	4.5	0.0	136.0	NaN	NaN	6.2	20.0	7.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0	
0.0	...	NaN	NaN	NaN	0.30	0.37	6.2	3.0	3.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	
0.0	...	NaN	NaN	NaN	0.38	0.11	6.9	30.0	3.0	0.0	0.0	0.0	NaN	NaN	NaN	0.0	
0.0	...	3.1	0.0	132.0	0.15	0.45	9.1	5.0	1.0	0.0	1.0	0.0	1.0	1.0	1.0	0.0	
0.0	...	4.1	0.0	140.0	1.13	0.60	9.6	18.0	2.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	
0.0	...	NaN	NaN	NaN	0.23	0.15	5.4	15.0	4.0	2.0	0.0	1.0	0.0	1.0	0.0	0.0	
0.0	...	3.4	0.0	132.0	1.05	0.37	5.0	27.0	2.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	
0.0	...	3.5	0.0	140.0	0.45	0.15	6.7	10.0	5.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	
0.0	...	4.0	1.0	150.0	0.45	0.22	3.9	10.0	3.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	
0.0	...	NaN	NaN	NaN	0.45	0.07	12.5	10.0	1.0	1.0	0.0	1.0	1.0	1.0	0.0	0.0	
0.0	...	3.8	0.0	138.0	0.30	0.15	4.4	19.0	2.0	1.0	0.0	1.0	1.0	0.0	0.0	1.0	
0.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	9.0	0.0	0.0	0.0	NaN	NaN	NaN	0.0	
0.0	...	4.2	0.0	130.0	0.15	0.15	7.3	24.0	9.0	0.0	0.0	0.0	NaN	NaN	NaN	0.0	

Dopo dell'applicazione del metodo:

GT_POST	...	K_BLOOD	GIPER_NA	NA_BLOOD	ALT_BLOOD	AST_BLOOD	L_BLOOD	ROE	TIME_B_S	R_AB_1_n	R_AB_2_n	R_AB_3_n	NA_KB	NOT_NA_KB	LID_KB	NITR_S	I
0.0	...	4.70	0.00	138.00	0.38	0.15	8.00	16.00	4.0	0.0	0.0	1.0	0.67	0.67	0.33	0.0	
0.0	...	3.50	0.00	132.00	0.38	0.18	7.80	3.00	2.0	0.0	0.0	0.0	1.00	0.00	1.00	0.0	
0.0	...	4.00	0.00	132.00	0.30	0.11	10.80	9.67	3.0	3.0	0.0	0.0	1.00	1.00	1.00	0.0	
0.0	...	3.90	0.00	146.00	0.75	0.37	6.80	10.00	2.0	0.0	0.0	1.0	0.33	1.00	0.00	0.0	
0.0	...	3.50	0.00	132.00	0.45	0.22	8.30	16.33	9.0	0.0	0.0	0.0	0.00	0.00	0.00	0.0	
0.0	...	3.70	0.00	133.33	0.45	0.22	7.20	2.00	2.0	0.0	0.0	0.0	0.00	1.00	0.00	0.0	
0.0	...	4.00	0.33	144.67	0.30	0.11	11.10	5.00	1.0	0.0	0.0	0.0	0.00	1.00	0.00	0.0	
0.0	...	4.50	0.00	136.00	0.52	0.25	6.20	20.00	7.0	3.0	0.0	0.0	0.00	0.00	0.00	0.0	
0.0	...	4.33	0.00	135.33	0.30	0.37	6.20	3.00	3.0	0.0	0.0	0.0	0.00	1.00	0.00	0.0	
0.0	...	4.10	0.00	133.33	0.38	0.11	6.90	30.00	3.0	0.0	0.0	0.0	0.67	0.67	0.33	0.0	
0.0	...	3.10	0.00	132.00	0.15	0.45	9.10	5.00	1.0	0.0	1.0	0.0	1.00	1.00	1.00	0.0	
0.0	...	4.10	0.00	140.00	1.13	0.60	9.60	18.00	2.0	0.0	0.0	0.0	0.00	1.00	0.00	0.0	
0.0	...	3.83	0.33	140.67	0.23	0.15	5.40	15.00	4.0	2.0	0.0	1.0	0.00	1.00	0.00	0.0	
0.0	...	3.40	0.00	132.00	1.05	0.37	5.00	27.00	2.0	0.0	0.0	0.0	1.00	1.00	0.00	0.0	
0.0	...	3.50	0.00	140.00	0.45	0.15	6.70	10.00	5.0	0.0	0.0	0.0	0.00	1.00	0.00	0.0	
0.0	...	4.00	1.00	150.00	0.45	0.22	3.90	10.00	3.0	0.0	0.0	0.0	1.00	1.00	0.00	0.0	
0.0	...	4.00	0.33	144.67	0.45	0.07	12.50	10.00	1.0	1.0	0.0	1.0	1.00	1.00	0.00	0.0	
0.0	...	3.80	0.00	138.00	0.30	0.15	4.40	19.00	2.0	1.0	0.0	1.0	1.00	0.00	0.00	1.0	
0.0	...	3.80	0.00	135.33	0.88	0.37	7.53	20.00	9.0	0.0	0.0	0.0	0.67	1.00	0.00	0.0	
0.0	...	4.20	0.00	130.00	0.15	0.15	7.30	24.00	9.0	0.0	0.0	0.0	0.67	1.00	0.33	0.0	

DATA SCALING

MYOCARDIAL INFARCTION-COMPLICATIONS

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

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

Al termine di questa fase di preprocessing si esportano i dati in formato CSV.

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

Si considerano gli algoritmi di clustering gerarchico di tipo Agglomerativo, occorre effettuare una ricerca degli iper-parametri per determinare:

- Il parametro di k
- Il metodo di linkage (Single, Average, Complete)

Per scegliere il modello si effettua una ricerca nello spazio dei modelli + spazio dei parametri

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 metrica 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

MYOCARDIAL INFARCTION-COMPLICATIONS

VALUTAZIONE DEI MODELLI

CRITERI ESTERNI

Per valutare i risultati del cluster scelto potresti 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.

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.

GINI E PURITY

Problema:

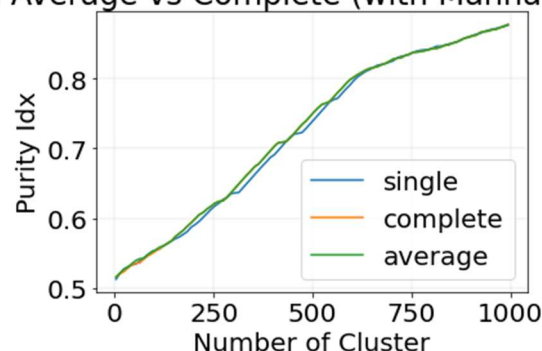
Le **variabili di output** a disposizione **non sono mutamente esclusive**, ma ogni osservazione è etichetta con più di una label.

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 (in base alle classi) tra loro.

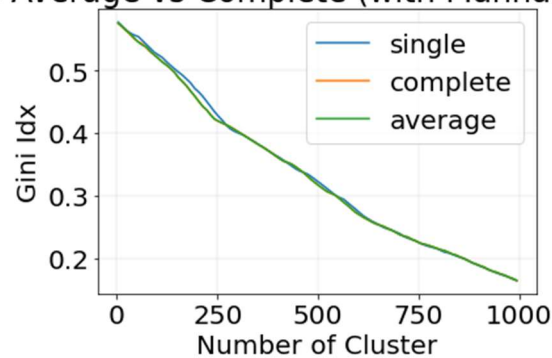
I criteri esterni Gini e Purity tenderanno a valutare meglio soluzioni con un numero di cluster elevato, sino ad arrivare a tanti cluster quante sono le osservazioni.

Single vs Average vs Complete (with Manhattan Distance)

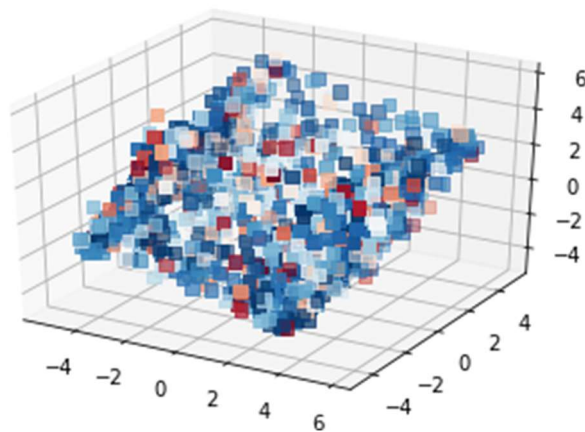


MYOCARDIAL INFARCTION-COMPLICATIONS

Single vs Average vs Complete (with Manhattan Distance)



Anche nel caso sia l'indice purity che gini mostrano risultati buoni solamente oltre ad un numero di cluster elevato dell'algoritmo.



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 e il cluster predetto è etichettato con il numero **7** allora in formato dummy sarà **0 0 0 0 0 0 1 0 0 0**.

MYOCARDIAL INFARCTION-COMPLICATIONS

Un criterio utile 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.

Dettagli sul criterio OC:

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

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

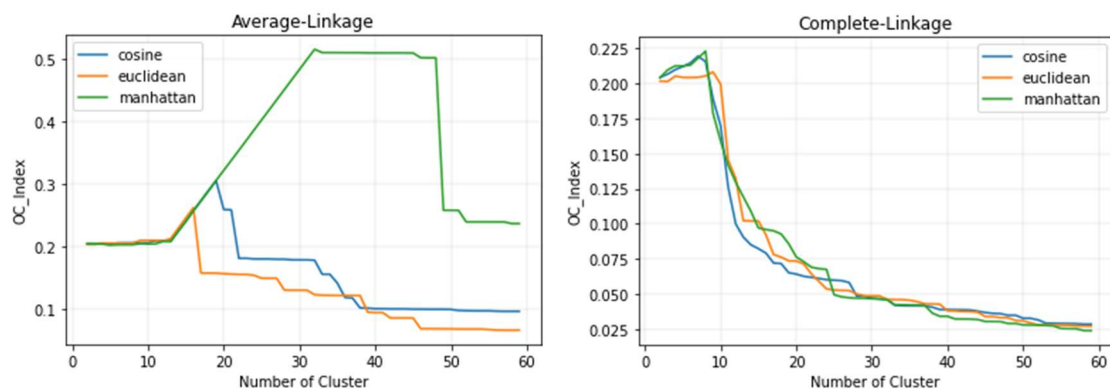
$$\tilde{t} = \frac{\sum_{i=1}^k \sum_{j=1}^{k'} \binom{|c_i \cap c'_j|}{2}}{\binom{N}{2} \frac{\max(n, n')}{N} \min(k, k')},$$

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

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

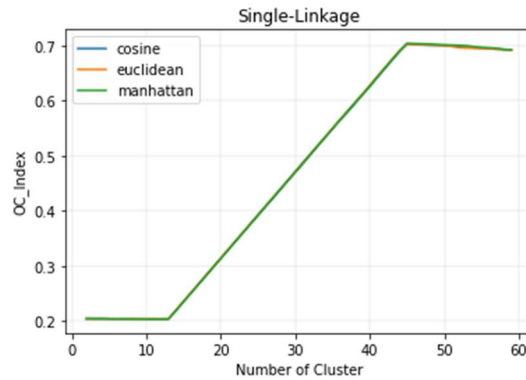
OC INDEX

Utilizzando come criterio esterno OC avremo i seguenti risultati.



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>

MYOCARDIAL INFARCTION-COMPLICATIONS



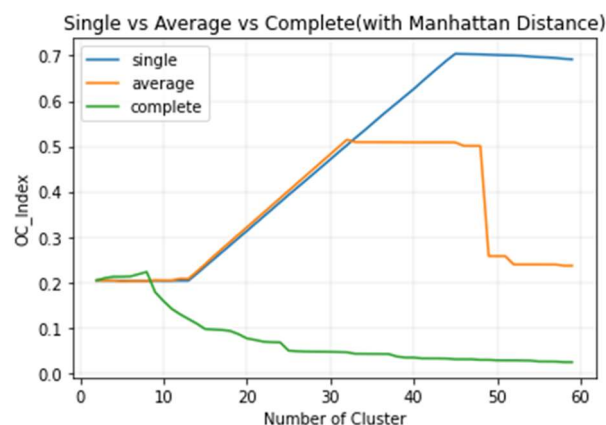
Il grafici mostrano le performance valutate con L'indice OC degli algoritmi di clustering agglomerativi Single-Linkage, Average-Linkage e Complete-Linkage al variare del numero K di cluster e della funzione di distanza utilizzata.

Nel caso del Single -Linkage tutte le distanza portano allo stesso risultato si hanno migliori risultati con un numero di cluster tra i 40 e 50.

Nel caso dell'Average-Linkage la distanza di Mahnathan permette di ottenere performance nettamente migliori delle altre distanza che infatti da un certo nuemro di clsuter (tra 15 e 20) tendono a scendere verso lo 0 mentre con la Mahnathan si arriva al "gomito" dopo i 30 cluster per poi descrescere.

Con il Complete-Linkage i risutlati non risultano tano diversi come nel caso dell'Average, ma anche in questo caso la distanza di Mhanathan sembra porta alla soluzione migliore.

A parità di funzioni di distanza occorre confrontare i vari algoritmi:



L'algoritmo Single-Linkage sembra quindi portare alla miglior soluzione.

```
C_0 size : 1525 labels : {"('A_V_BLOK', 'DRESSLER')": 1, "('A_V_BLOK', 'LET_IS')": 2, "('A_V_BLOK', 'OTEK_LANC')": 1, "('A_V_BLOK', 'OTEK_LANC', 'LET_IS')": 1, "('A_V_BLOK', 'OTEK_LANC', 'REC_IM')": 2, "('A_V_BLOK', 'OTEK_LANC', 'ZSN', 'REC_IM')": 1, "('A_V_BLOK',
```

MYOCARDIAL INFARCTION-COMPLICATIONS

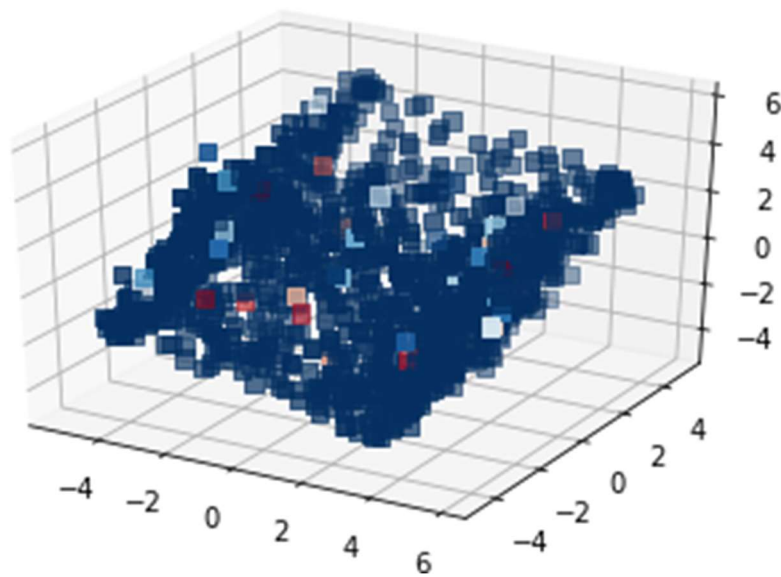
```
'RAZRIV', 'LET_IS')": 3, "('A_V_BLOK', 'REC_IM')": 2, "('A_V_BLOK',
'REC_IM', 'LET_IS')": 1, "('A_V_BLOK', 'ZSN')": 7, "('A_V_BLOK',)":
15, "('DRESSLER', 'P_IM_STEN')": 1, "('DRESSLER', 'ZSN')": 14,
"('DRESSLER', 'ZSN', 'REC_IM')": 2, "('DRESSLER',)": 38,
"('FIBR_JELUD', 'A_V_BLOK')": 1, "('FIBR_JELUD', 'A_V_BLOK',
'LET_IS')": 1, "('FIBR_JELUD', 'A_V_BLOK', 'RAZRIV', 'LET_IS')": 1,
"('FIBR_JELUD', 'A_V_BLOK', 'ZSN')": 2, "('FIBR_JELUD', 'LET_IS')": 6,
"('FIBR_JELUD', 'OTEK_LANC', 'DRESSLER', 'REC_IM')": 1,
"('FIBR_JELUD', 'OTEK_LANC', 'DRESSLER', 'ZSN')": 1, "('FIBR_JELUD',
'P_IM_STEN')": 1, "('FIBR_JELUD', 'RAZRIV', 'LET_IS')": 3,
"('FIBR_JELUD', 'REC_IM')": 1, "('FIBR_JELUD', 'REC_IM',
'P_IM_STEN')": 1, "('FIBR_JELUD', 'ZSN')": 2, "('FIBR_JELUD', 'ZSN',
'LET_IS')": 1, "('FIBR_JELUD',)": 21, "('FIBR_PREDS', 'A_V_BLOK')": 2,
"('FIBR_PREDS', 'A_V_BLOK', 'OTEK_LANC', 'LET_IS')": 1,
"('FIBR_PREDS', 'A_V_BLOK', 'RAZRIV', 'LET_IS')": 1, "('FIBR_PREDS',
'A_V_BLOK', 'ZSN')": 2, "('FIBR_PREDS', 'DRESSLER', 'ZSN')": 1,
"('FIBR_PREDS', 'DRESSLER', 'ZSN', 'LET_IS')": 1, "('FIBR_PREDS',
'DRESSLER', 'ZSN', 'P_IM_STEN')": 1, "('FIBR_PREDS', 'FIBR_JELUD')":
3, "('FIBR_PREDS', 'FIBR_JELUD', 'LET_IS')": 1, "('FIBR_PREDS',
'FIBR_JELUD', 'OTEK_LANC')": 1, "('FIBR_PREDS', 'FIBR_JELUD',
'REC_IM', 'LET_IS')": 1, "('FIBR_PREDS', 'FIBR_JELUD', 'ZSN')": 1,
"('FIBR_PREDS', 'FIBR_JELUD', 'ZSN', 'REC_IM')": 2, "('FIBR_PREDS',
'JELUD_TAH', 'A_V_BLOK', 'LET_IS')": 1, "('FIBR_PREDS', 'JELUD_TAH',
'FIBR_JELUD', 'DRESSLER', 'ZSN')": 1, "('FIBR_PREDS', 'JELUD_TAH',
'FIBR_JELUD', 'OTEK_LANC')": 1, "('FIBR_PREDS', 'JELUD_TAH',
'FIBR_JELUD', 'REC_IM', 'LET_IS')": 1, "('FIBR_PREDS', 'JELUD_TAH',
'OTEK_LANC', 'ZSN', 'REC_IM')": 1, "('FIBR_PREDS', 'JELUD_TAH',
'RAZRIV', 'LET_IS')": 1, "('FIBR_PREDS', 'LET_IS')": 4,
"('FIBR_PREDS', 'OTEK_LANC')": 5, "('FIBR_PREDS', 'OTEK_LANC',
'ZSN')": 3, "('FIBR_PREDS', 'OTEK_LANC', 'ZSN', 'LET_IS')": 1,
"('FIBR_PREDS', 'PREDS_TAH')": 2, "('FIBR_PREDS', 'PREDS_TAH',
'DRESSLER', 'ZSN')": 1, "('FIBR_PREDS', 'PREDS_TAH', 'LET_IS')": 2,
"('FIBR_PREDS', 'PREDS_TAH', 'OTEK_LANC', 'ZSN', 'LET_IS')": 1,
"('FIBR_PREDS', 'P_IM_STEN')": 3, "('FIBR_PREDS', 'RAZRIV',
'DRESSLER', 'LET_IS')": 1, "('FIBR_PREDS', 'RAZRIV', 'LET_IS')": 1,
"('FIBR_PREDS', 'RAZRIV', 'REC_IM', 'LET_IS')": 1, "('FIBR_PREDS',
'REC_IM')": 3, "('FIBR_PREDS', 'REC_IM', 'LET_IS')": 5,
"('FIBR_PREDS', 'REC_IM', 'P_IM_STEN')": 1, "('FIBR_PREDS', 'ZSN')":
27, "('FIBR_PREDS', 'ZSN', 'LET_IS')": 1, "('FIBR_PREDS', 'ZSN',
'REC_IM')": 3, "('FIBR_PREDS', 'ZSN', 'REC_IM', 'LET_IS')": 1,
"('FIBR_PREDS',)": 58, "('JELUD_TAH', 'A_V_BLOK')": 1, "('JELUD_TAH',
'A_V_BLOK', 'LET_IS')": 1, "('JELUD_TAH', 'DRESSLER', 'REC_IM',
'P_IM_STEN')": 1, "('JELUD_TAH', 'FIBR_JELUD')": 1, "('JELUD_TAH',
'FIBR_JELUD', 'A_V_BLOK', 'ZSN', 'LET_IS')": 1, "('JELUD_TAH',
'FIBR_JELUD', 'DRESSLER', 'REC_IM')": 1, "('JELUD_TAH', 'FIBR_JELUD',
'OTEK_LANC')": 1, "('JELUD_TAH', 'FIBR_JELUD', 'ZSN', 'REC_IM')": 1,
"('JELUD_TAH', 'OTEK_LANC', 'DRESSLER', 'ZSN')": 1, "('JELUD_TAH',
'OTEK_LANC', 'REC_IM', 'LET_IS')": 1, "('JELUD_TAH', 'REC_IM',
'LET_IS')": 1, "('JELUD_TAH', 'ZSN')": 5, "('JELUD_TAH', 'ZSN',
'REC_IM')": 1, "('JELUD_TAH',)": 14, "('LET_IS',)": 72,
"('NO_COMP',)": 624, "('OTEK_LANC', 'DRESSLER')": 2, "('OTEK_LANC',
'DRESSLER', 'ZSN', 'REC_IM')": 1, "('OTEK_LANC', 'DRESSLER', 'ZSN',
'REC_IM', 'LET_IS')": 1, "('OTEK_LANC', 'LET_IS')": 20, "('OTEK_LANC',
'P_IM_STEN')": 1, "('OTEK_LANC', 'RAZRIV', 'REC_IM', 'LET_IS')": 1,
"('OTEK_LANC', 'REC_IM')": 11, "('OTEK_LANC', 'REC_IM', 'LET_IS')": 1,
"('OTEK_LANC', 'REC_IM', 'P_IM_STEN')": 1, "('OTEK_LANC', 'ZSN')": 24,
"('OTEK_LANC', 'ZSN', 'LET_IS')": 3, "('OTEK_LANC', 'ZSN',
'P_IM_STEN')": 2, "('OTEK_LANC', 'ZSN', 'REC_IM')": 7, "('OTEK_LANC',
'ZSN', 'REC_IM', 'LET_IS')": 3, "('OTEK_LANC',)": 34, "('PREDS_TAH',
'JELUD_TAH')": 1, "('PREDS_TAH', 'REC_IM')": 1, "('PREDS_TAH',
'REC_IM', 'P_IM_STEN')": 1, "('PREDS_TAH', 'ZSN')": 2,
"('PREDS_TAH',)": 5, "('P_IM_STEN', 'LET_IS')": 1, "('P_IM_STEN',)":
```

MYOCARDIAL INFARCTION-COMPLICATIONS

```
88, "('RAZRIV', 'LET_IS')": 32, "('RAZRIV', 'REC_IM', 'LET_IS')": 1,
"('RAZRIV', 'ZSN', 'LET_IS')": 1, "('REC_IM', 'LET_IS')": 15,
"('REC_IM', 'P_IM_STEN')": 6, "('REC_IM',)": 30, "('ZSN', 'LET_IS')":
8, "('ZSN', 'P_IM_STEN')": 14, "('ZSN', 'REC_IM')": 9, "('ZSN',
'REC_IM', 'LET_IS')": 4, "('ZSN', 'REC_IM', 'P_IM_STEN')": 5,
"('ZSN',)": 180}
C_1 size : 2 labels : {"('LET_IS',)": 2}
C_2 size : 1 labels : {"('RAZRIV', 'LET_IS')": 1}
C_3 size : 1 labels : {"('LET_IS',)": 1}
C_4 size : 1 labels : {"('LET_IS',)": 1}
C_5 size : 1 labels : {"('LET_IS',)": 1}
C_6 size : 1 labels : {"('OTEK_LANC', 'ZSN', 'LET_IS')": 1}
C_7 size : 1 labels : {"('OTEK_LANC', 'ZSN', 'LET_IS')": 1}
C_8 size : 1 labels : {"('LET_IS',)": 1}
C_9 size : 1 labels : {"('LET_IS',)": 1}
C_10 size : 1 labels : {"('ZSN', 'REC_IM', 'LET_IS')": 1}
C_11 size : 1 labels : {"('LET_IS',)": 1}
C_12 size : 1 labels : {"('FIBR_PREDS', 'RAZRIV', 'ZSN', 'LET_IS')":
1}
C_13 size : 1 labels : {"('FIBR_PREDS', 'RAZRIV', 'ZSN', 'LET_IS')":
1}
C_14 size : 1 labels : {"('FIBR_JELUD', 'ZSN', 'REC_IM', 'LET_IS')":
1}
C_15 size : 1 labels : {"('FIBR_JELUD', 'LET_IS')": 1}
C_16 size : 1 labels : {"('LET_IS',)": 1}
C_17 size : 1 labels : {"('REC_IM', 'LET_IS')": 1}
C_18 size : 1 labels : {"('RAZRIV', 'LET_IS')": 1}
C_19 size : 1 labels : {"('LET_IS',)": 1}
C_20 size : 1 labels : {"('LET_IS',)": 1}
C_21 size : 1 labels : {"('LET_IS',)": 1}
C_22 size : 1 labels : {"('JELUD_TAH', 'FIBR_JELUD', 'LET_IS')": 1}
C_23 size : 1 labels : {"('ZSN',)": 1}
C_24 size : 1 labels : {"('OTEK_LANC', 'ZSN', 'REC_IM', 'LET_IS')": 1}
C_25 size : 1 labels : {"('NO_COMP',)": 1}
C_26 size : 1 labels : {"('LET_IS',)": 1}
C_27 size : 1 labels : {"('FIBR_PREDS', 'DRESSLER', 'P_IM_STEN')": 1}
C_28 size : 1 labels : {"('FIBR_PREDS', 'OTEK_LANC', 'ZSN', 'REC_IM',
'LET_IS')": 1}
C_29 size : 1 labels : {"('OTEK_LANC', 'RAZRIV', 'ZSN', 'LET_IS')": 1}
C_30 size : 1 labels : {"('ZSN', 'LET_IS')": 1}
C_31 size : 1 labels : {"('NO_COMP',)": 1}
C_32 size : 1 labels : {"('ZSN',)": 1}
C_33 size : 1 labels : {"('OTEK_LANC', 'LET_IS')": 1}
C_34 size : 1 labels : {"('FIBR_PREDS', 'FIBR_JELUD', 'ZSN')": 1}
C_35 size : 1 labels : {"('NO_COMP',)": 1}
C_36 size : 1 labels : {"('REC_IM', 'LET_IS')": 1}
C_37 size : 1 labels : {"('NO_COMP',)": 1}
C_38 size : 1 labels : {"('A_V_BLOK', 'LET_IS')": 1}
C_39 size : 1 labels : {"('RAZRIV', 'LET_IS')": 1}
C_40 size : 1 labels : {"('LET_IS',)": 1}
C_41 size : 1 labels : {"('FIBR_PREDS', 'LET_IS')": 1}
C_42 size : 1 labels : {"('LET_IS',)": 1}
C_43 size : 1 labels : {"('FIBR_PREDS', 'PREDS_TAH', 'FIBR_JELUD',
'LET_IS')": 1}
C_44 size : 1 labels : {"('JELUD_TAH', 'OTEK_LANC')": 1}
```

Proiettando le tre componenti principali della PCA applicata dei dati è possibile visualizzare il clustering effettuato.

MYOCARDIAL INFARCTION-COMPLICATIONS

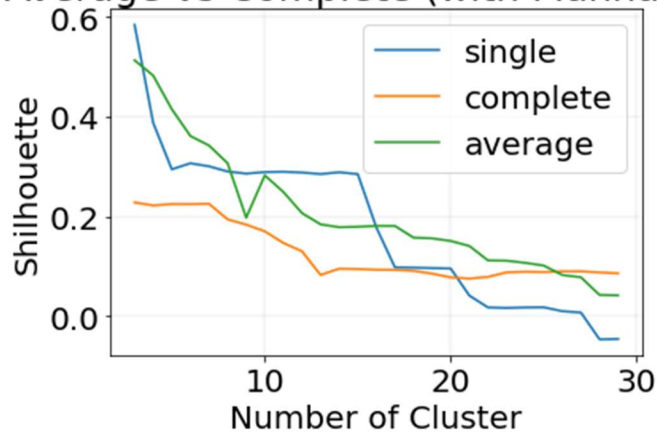


Si nota la prevalenza di un cluster di dimensione elevata rispetto agli altri di dimensione notevolmente inferiore.

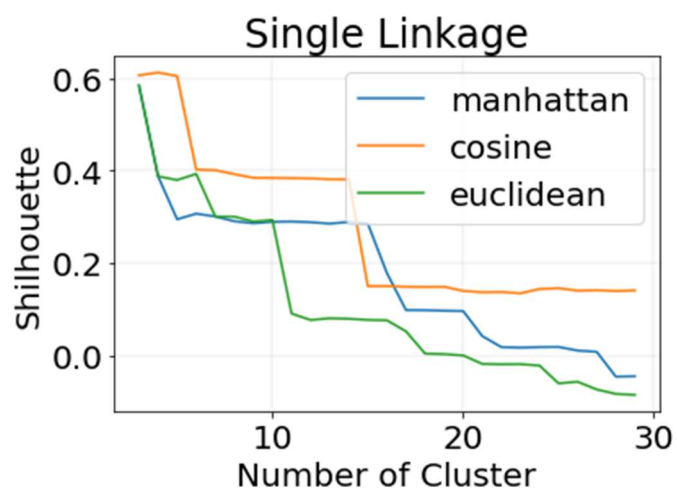
CRITERI INTERNI

INDICE DI SHILOETTE

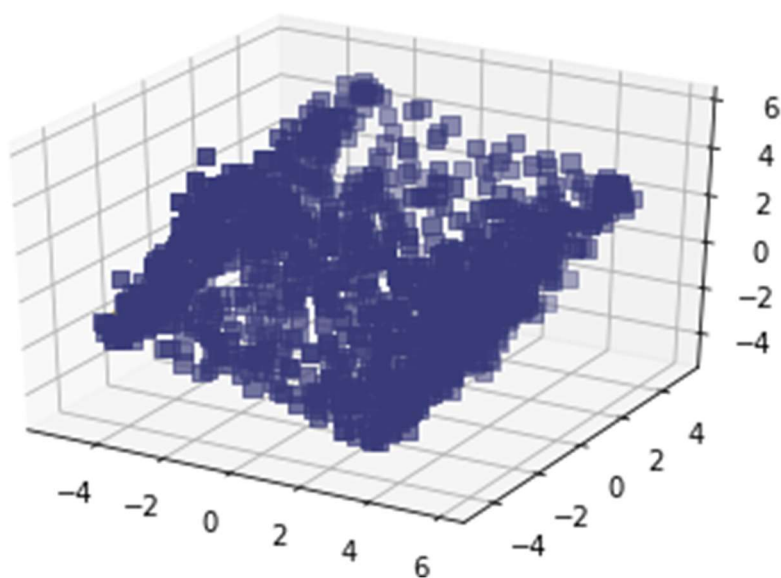
Single vs Average vs Complete (with Manhattan Distance)



MYOCARDIAL INFARCTION-COMPLICATIONS



Il numero di cluster ottimale per questo criterio è tra 2 e 4, con il single linkage e la distanza del coseno, la soluzione con $k = 2$ è visualizzata come segue:



MYOCARDIAL INFARCTION-COMPLICATIONS

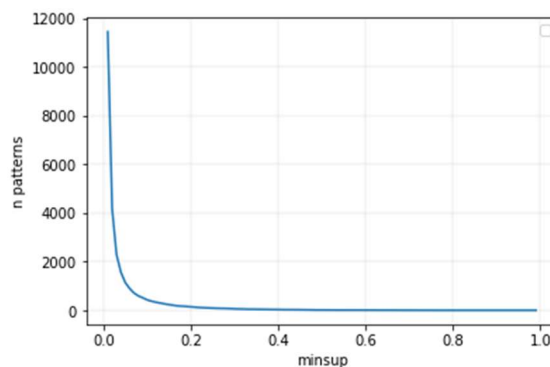
ESTRAZIONE DI REGOLE DI ASSOCIAZIONE

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.

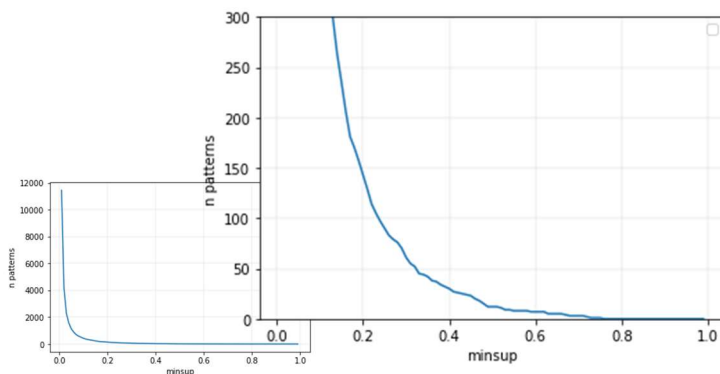
Il dataset ristretto alle soli variabili binari viene quindi considerato come un insieme di transazioni.

	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
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	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
2	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
3	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
4	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
...
1662	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
1663	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
1664	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
1665	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
1666	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

1667 rows x 21 columns



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 40-60% i pattern tendono a diminuire drasticamente sino ad arrivare a non scoprire nessun pattern.

MYOCARDIAL INFARCTION-COMPLICATIONS

Si optato quindi per una percentuale di soglia di minsup del 15% 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

MYOCARDIAL INFARCTION-COMPLICATIONS

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
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', 'GEPARS_n', '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},
```


MYOCARDIAL INFARCTION-COMPLICATIONS

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

MYOCARDIAL INFARCTION-COMPLICATIONS

```
{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

MYOCARDIAL INFARCTION-COMPLICATIONS

Analisi dei cluster:

I risultati riportati dai criteri esterni ed interni sono opposti: da una parte i criteri interni valutano meglio soluzioni con un minor numero di cluster, dall'altra i criteri esterni spingono verso un numero di cluster elevato. Se è visto che utilizzando come criterio esterno OC si ottiene una soluzione con un numero intermedio di cluster, tale soluzione però consiste in un grande cluster e tanti piccoli cluster di dimensioni ridotte. Da questi risultati si arriva alla conclusione che per il numero di osservazioni del dataset disponibili i cluster naturali non sembrano in qualche modo correlate alle etichette del dataset, è quindi difficile interpretare o dare una caratterizzazione dei gruppi creati, poiché ogni cluster contiene un insieme molto eterogeneo di osservazioni.

Regole di associazione estratte:

L'algoritmo Apriori ha portato a delle informazioni interessanti tra le le regole di associazione 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

{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

MYOCARDIAL INFARCTION-COMPLICATIONS