MYOCARDIAL INFARCTION-COMPLICATIONS ANALYSIS

Data mining : Caso di Studio

Vito Simone Lacatena

2021

# 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. Associtation 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 complicaziazioni 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

|  |  |  |
| --- | --- | --- |
| **Nome** | **Descrizione** | **Diagramma** |
| FIBR\_PREDS | Atrial fibrillation |  |
| PREDS\_TAH | Supraventricular tachycardia |  |
| JELUD\_TAH | Ventricular tachycardia |  |
| FIBR\_JELUD | Ventricular fibrillation |  |
| A\_V\_BLOK) | Third-degree AV block |  |

|  |  |  |
| --- | --- | --- |
| **Nome** | **Descrizione** | **Diagramma** |
| OTEK\_LANC | Pulmonary edema |  |
| RAZRIV | Myocardial rupture |  |
| DRESSLER | Dressler syndrome |  |
| ZSN | Chronic heart failure |  |
| REC\_IM | Relapse of the myocardial infarction |  |

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:

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | FIBR\_PREDS | PREDS\_TAH | JELUD\_TAH | FIBR\_JELUD | A\_V\_BLOK | OTEK\_LAN | C    RAZRIV | DRESSLER | ZSN | REC\_IM | P\_IM\_STEN | LET\_IS |
| 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 | FIBR\_PREDS | PREDS\_TAH | JELUD\_TAH | FIBR\_JELUD | A\_V\_BLOK | OTEK\_LAN | C    RAZRIV | DRESSLER | ZSN | REC\_IM | P\_IM\_STEN | LET\_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 |
| 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 |  |  |  |  |  |  |  |  |
| 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 | FIBR\_PREDS | PREDS\_TAH | JELUD\_TAH | FIBR\_JELUD | A\_V\_BLOK | OTEK\_LAN | C    RAZRIV | DRESSLER | ZSN | REC\_IM | P\_IM\_STEN | LET\_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 | FIBR\_PREDS | PREDS\_TAH | JELUD\_TAH | FIBR\_JELUD | A\_V\_BLOK | OTEK\_LAN | C    RAZRIV | DRESSLER | ZSN | REC\_IM | P\_IM\_STEN | LET\_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 |  |  |  |
| 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 |  |  |  |  | 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 AlAT 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

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Name** | **Description** | **1** | **0** | **Missings** | **Pie** |
| SEX | Genere del pazione 0:Donna 1:Uomo | 62.65% | 37.35% | 0.0% |  |
| IBS\_NASL | Ereditarietà su CHD. 0: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** | **0** | **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% |  |
| zab\_leg\_04 | Asma bronchiale nell'anamnesi | 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% |  |
| 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% |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Name** | **Description** | **1** | **0** | **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 Сеliasum 750k IU | 0.76% | 98.65% | 0.59% |  |
| fibr\_ter\_02 | Terapia fibrinolitica con Сеliasum 1m IU | 0.94% | 98.47% | 0.59% |  |
| fibr\_ter\_03 | Terapia fibrinolitica con Сеliasum 3m IU | 4.0% | 95.41% | 0.59% |  |
| fibr\_ter\_05 | Terapia fibrinolitica con Streptase | 0.24% | 99.18% | 0.59% |  |
| fibr\_ter\_06 | Terapia fibrinolitica con Сеliasum 500k IU | 0.53% | 98.88% | 0.59% |  |
| fibr\_ter\_07 | Terapia fibrinolitica con Сеliasum 250k IU | 0.35% | 99.06% | 0.59% |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Name** | **Description** | **1** | **0** | **Missings** | **Pie** |
| fibr\_ter\_08 | 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 а 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% |  |
| 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

|  |  |  |  |
| --- | --- | --- | --- |
| **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),   1. (da 6-10 anni), 2. (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 IIА (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) |  |
| 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: medo di 2 ore  2: 2-4 ore  3: 4-6 ore  4: 6-8 ore   1. : 8-12 ore 2. : 12-24 ore 3. : 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) |  |

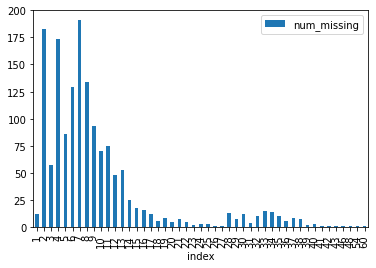
## Data Preparation

## **Problemi Riscontrati**

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

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:



**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%

SEX - 0%

FIBR\_PREDS - 0%

PREDS\_TAH - 0%

JELUD\_TAH - 0%

FIBR\_JELUD - 0%

A\_V\_BLOK - 0%

OTEK\_LANC - 0%

RAZRIV - 0%

DRESSLER - 0%

ZSN - 0%

REC\_IM - 0%

P\_IM\_STEN - 0%

LET\_IS - 0%

IM\_PG\_P - 0%

INF\_ANAM - 0%

NA\_R\_1\_n - 0%

zab\_leg\_01 - 0%

zab\_leg\_02 - 0%

zab\_leg\_03 - 0%

zab\_leg\_04 - 0%

zab\_leg\_06 - 0%

AGE - 1%

SIM\_GIPERT - 1%

NITR\_S - 1%

NOT\_NA\_1\_n - 1%

GB - 1%

fibr\_ter\_01 - 1%

fibr\_ter\_02 - 1%

fibr\_ter\_03 - 1%

fibr\_ter\_05 - 1%

fibr\_ter\_06 - 1%

fibr\_ter\_07 - 1%

fibr\_ter\_08 - 1%

endocr\_01 - 1%

endocr\_02 - 1%

endocr\_03 - 1%

O\_L\_POST - 1%

SVT\_POST - 1%

GT\_POST - 1%

FIB\_G\_POST - 1%

LID\_S\_n - 1%

B\_BLOK\_S\_n - 1%

MP\_TP\_POST - 1%

ANT\_CA\_S\_n - 1%

K\_SH\_POST - 1%

R\_AB\_1\_n - 1%

TIKL\_S\_n - 1%

TRENT\_S\_n - 1%

np\_01 - 1%

np\_04 - 1%

np\_05 - 1%

np\_07 - 1%

np\_08 - 1%

np\_09 - 1%

np\_10 - 1%

GEPAR\_S\_n - 1%

ASP\_S\_n - 1%

nr\_11 - 1%

nr\_01 - 1%

nr\_02 - 1%

nr\_03 - 1%

nr\_04 - 1%

nr\_07 - 1%

nr\_08 - 1%

IBS\_POST - 2%

ZSN\_A - 3%

FK\_STENOK - 4%

post\_im - 4%

lat\_im - 4%

ant\_im - 5%

inf\_im - 5%

STENOK\_AN - 6%

R\_AB\_2\_n - 6%

NA\_R\_2\_n - 6%

NOT\_NA\_2\_n - 6%

n\_p\_ecg\_p\_01 - 7%

n\_p\_ecg\_p\_03 - 7%

n\_p\_ecg\_p\_04 - 7%

n\_p\_ecg\_p\_05 - 7%

n\_p\_ecg\_p\_06 - 7%

n\_p\_ecg\_p\_07 - 7%

n\_p\_ecg\_p\_08 - 7%

n\_p\_ecg\_p\_09 - 7%

n\_p\_ecg\_p\_10 - 7%

n\_p\_ecg\_p\_11 - 7%

n\_p\_ecg\_p\_12 - 7%

n\_r\_ecg\_p\_01 - 7%

n\_r\_ecg\_p\_02 - 7%

n\_r\_ecg\_p\_03 - 7%

n\_r\_ecg\_p\_04 - 7%

n\_r\_ecg\_p\_05 - 7%

n\_r\_ecg\_p\_06 - 7%

n\_r\_ecg\_p\_08 - 7%

n\_r\_ecg\_p\_09 - 7%

n\_r\_ecg\_p\_10 - 7%

R\_AB\_3\_n - 7%

NA\_R\_3\_n - 7%

NOT\_NA\_3\_n - 7%

TIME\_B\_S - 7%

L\_BLOOD - 8%

ritm\_ecg\_p\_01 - 9%

ritm\_ecg\_p\_02 - 9%

ritm\_ecg\_p\_04 - 9%

ritm\_ecg\_p\_06 - 9%

ritm\_ecg\_p\_07 - 9%

ritm\_ecg\_p\_08 - 9%

ROE - 12%

DLIT\_AG - 15%

S\_AD\_ORIT - 17%

D\_AD\_ORIT - 17%

ALT\_BLOOD - 17%

AST\_BLOOD - 17%

GIPO\_K - 21%

K\_BLOOD - 22%

GIPER\_NA - 22%

NA\_BLOOD - 22%

NA\_KB - 39%

LID\_KB - 40%

**NOT\_NA\_KB - 41%**

**S\_AD\_KBRIG - 63%**

**D\_AD\_KBRIG - 63%**

**IBS\_NASL - 96%**

**KFK\_BLOOD - 100%**

**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 percentaule 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)

Drop di osservazioni

Per eliminare le osservazioni senza troppe perdite, occorre stabile 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 di dei K vicini più vicini trovati nel dataset. Due osservazioni sono vicine se le features (con valori non mancanti) sono vicine,.

Si è stato settato 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

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.

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:

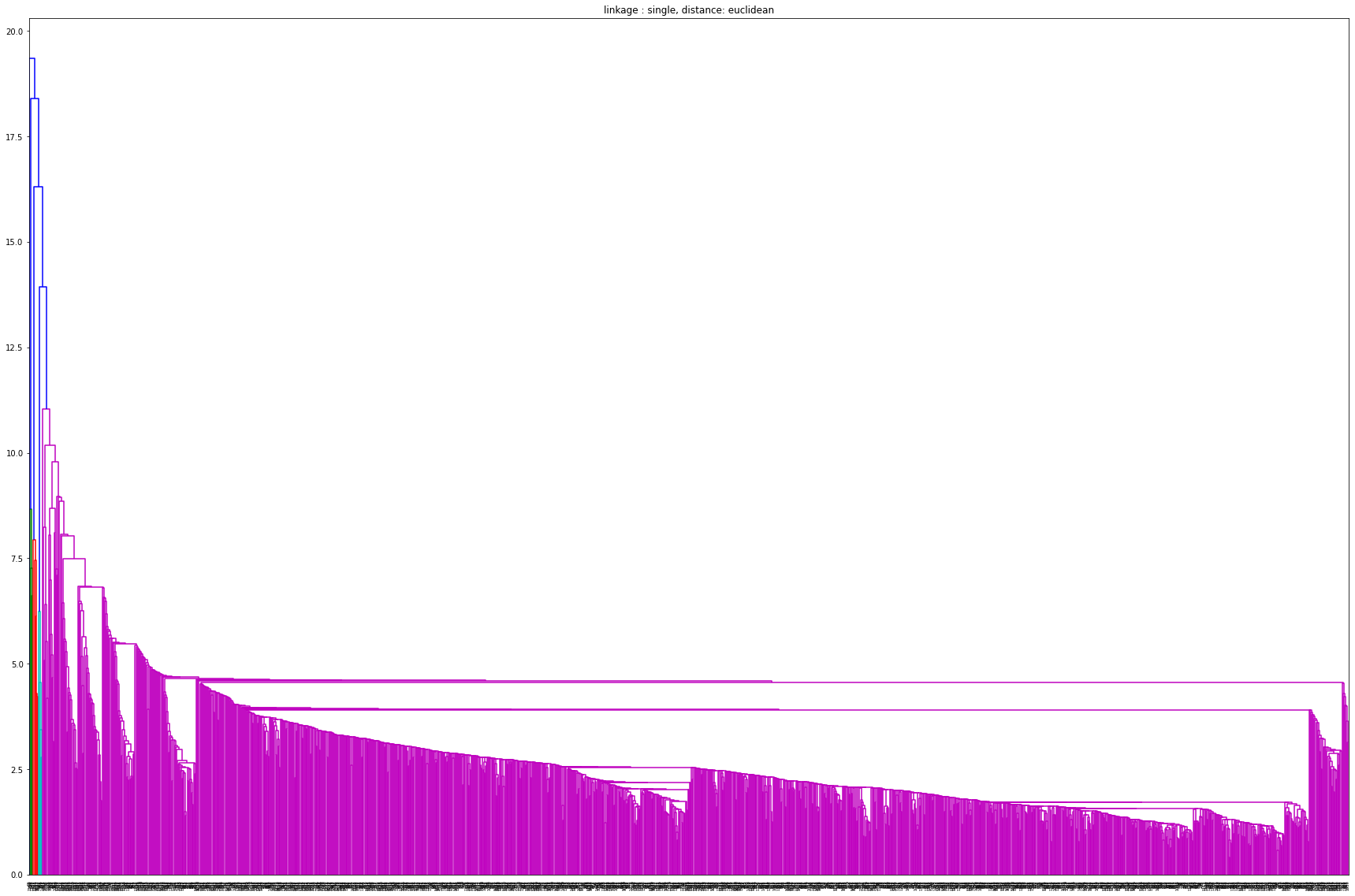
* Il parametro di k : **2 – n**

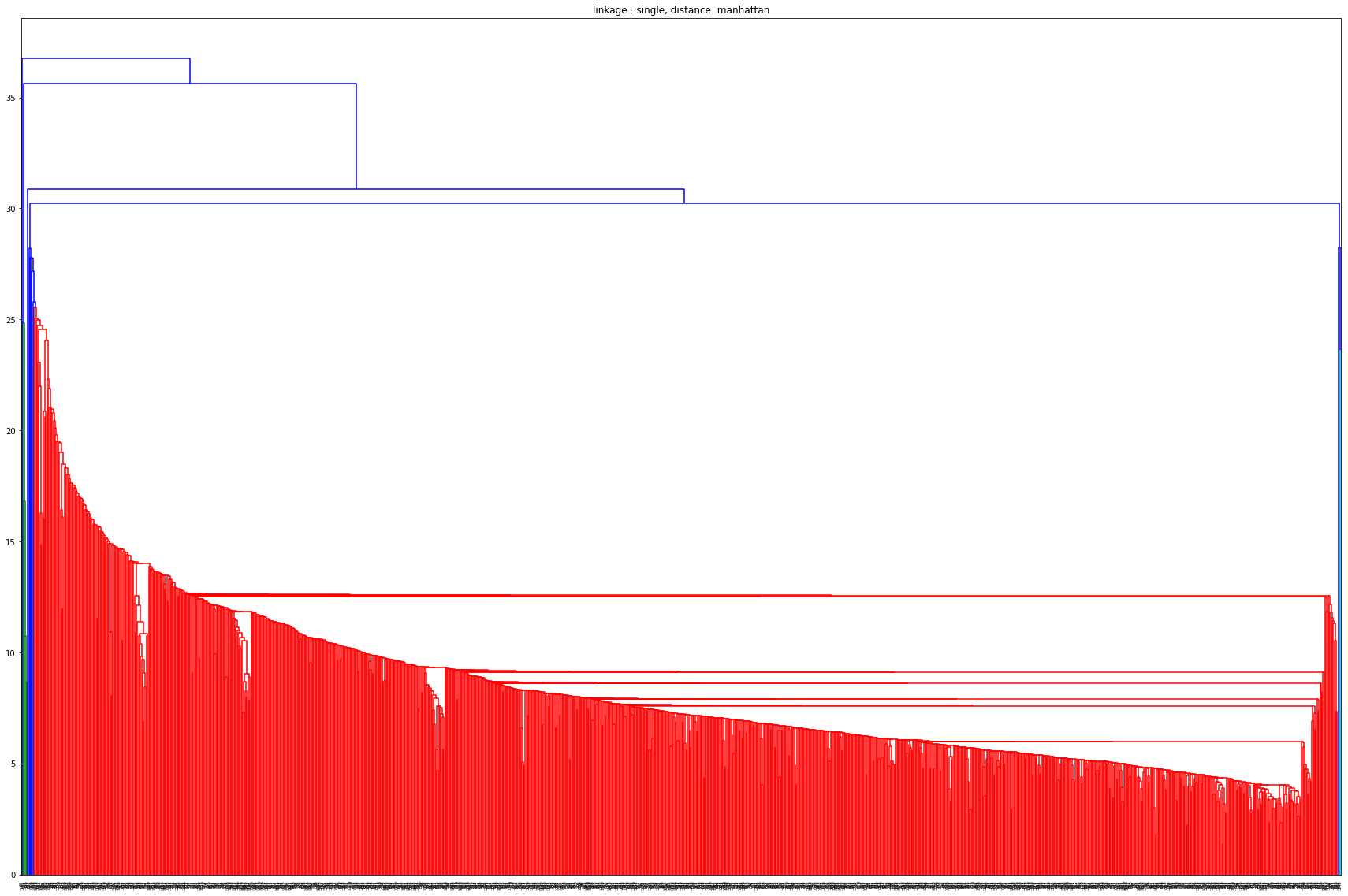
Il metodo di linkage : **Single, Average,Complete ,Ward**

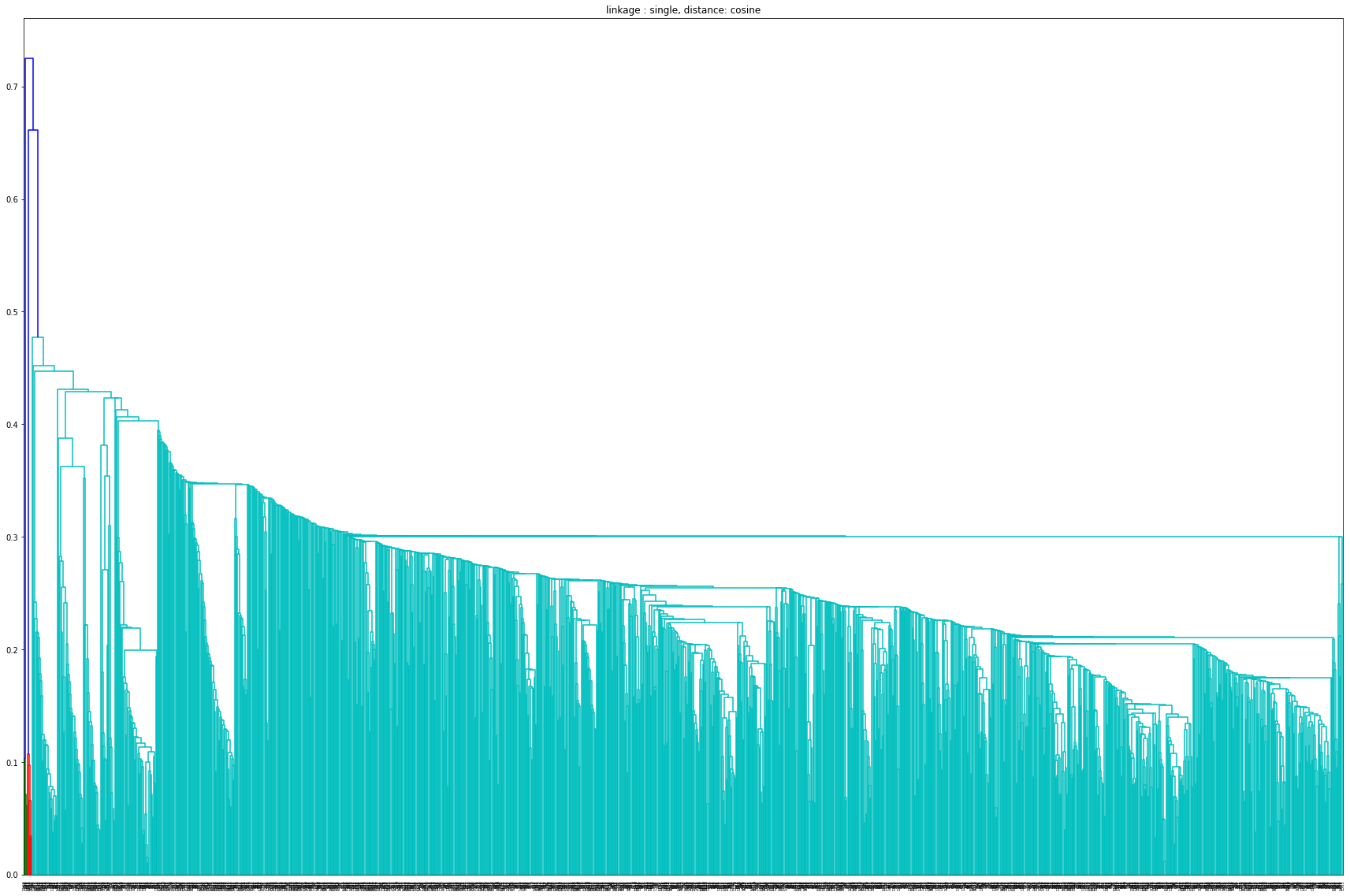
* La misura di distanza : **Euclidea, Coseno, Manhattan**

Dendogrammi:

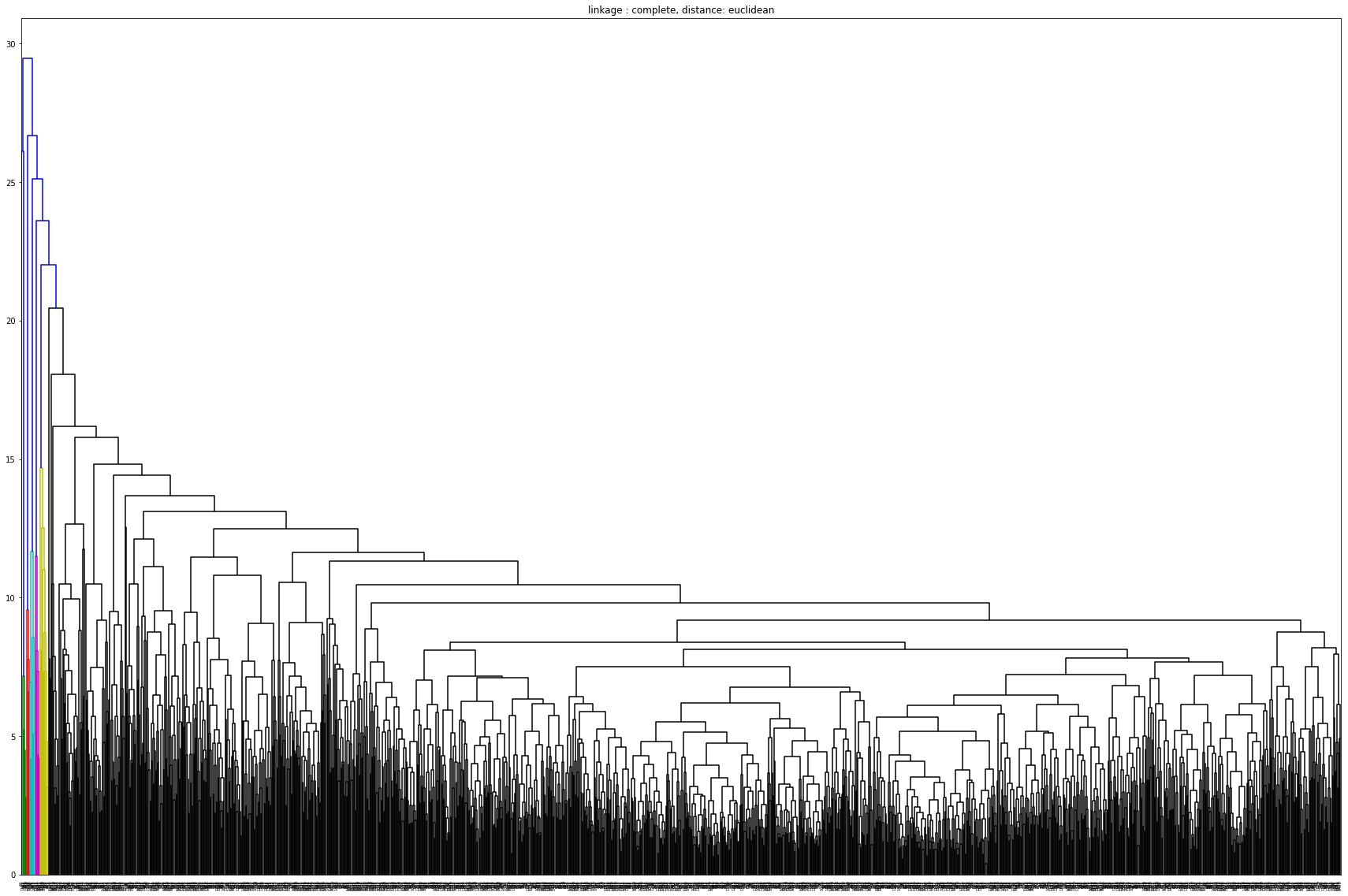
Single Linkage

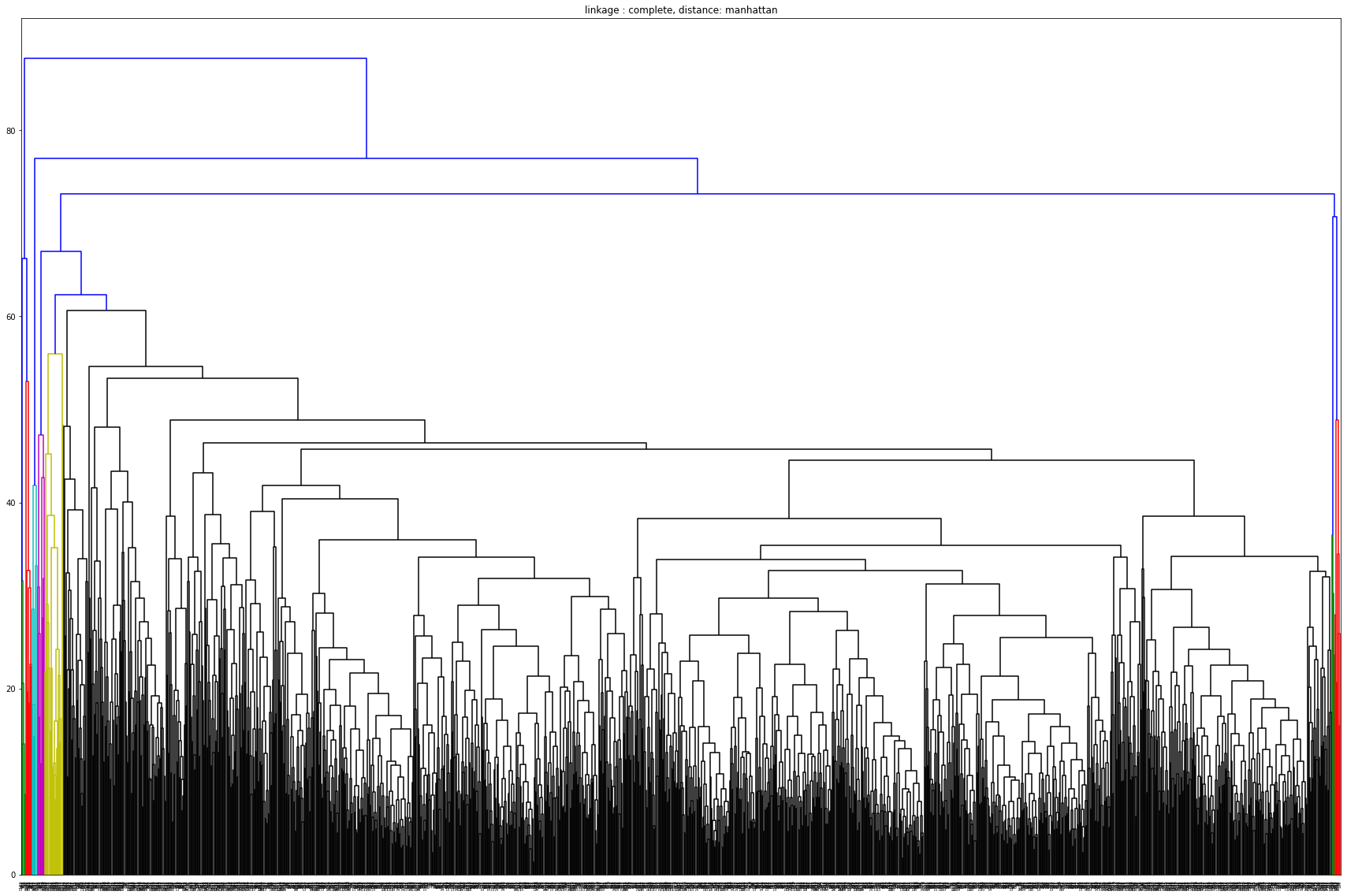
****

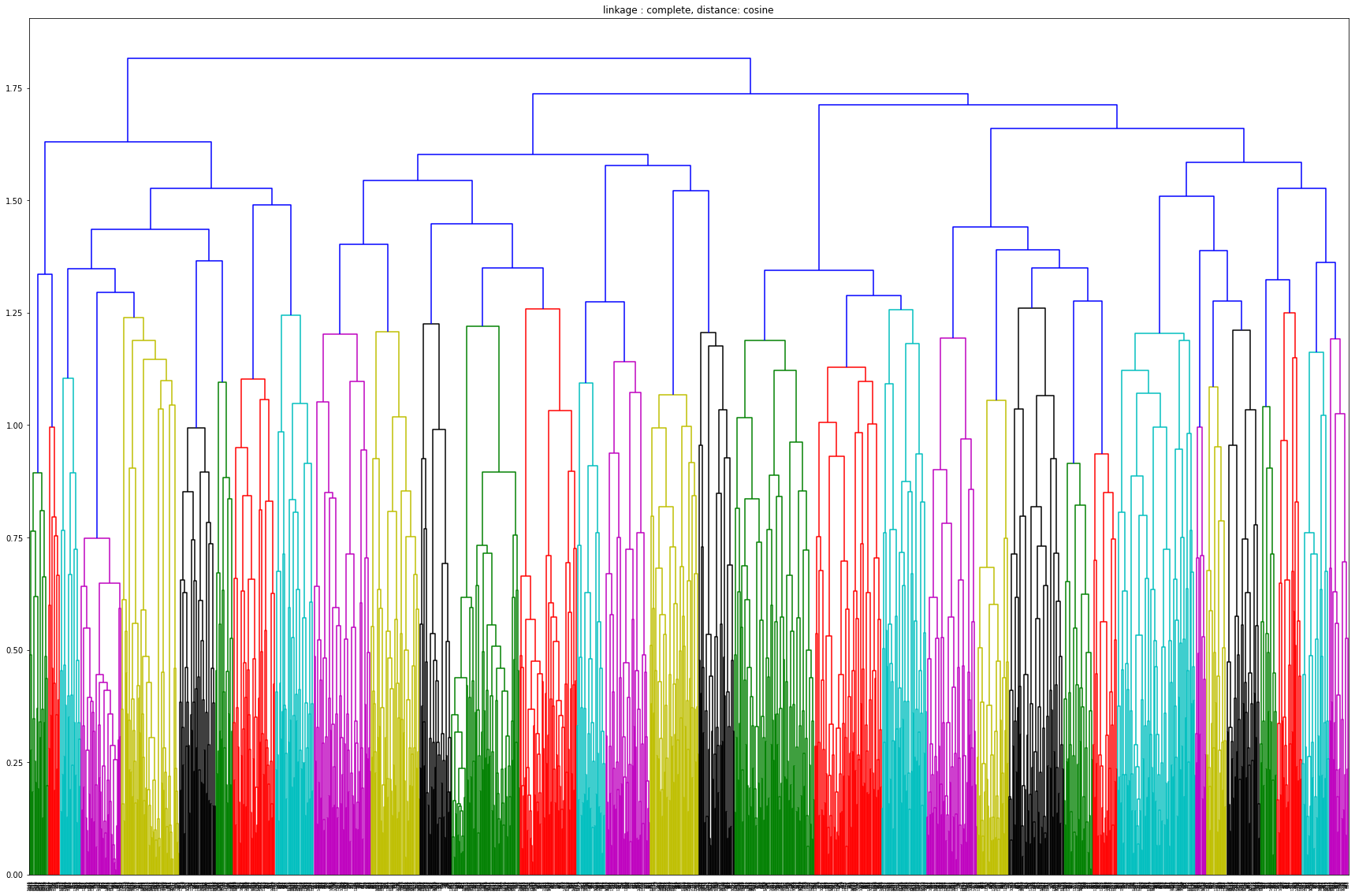
****

****

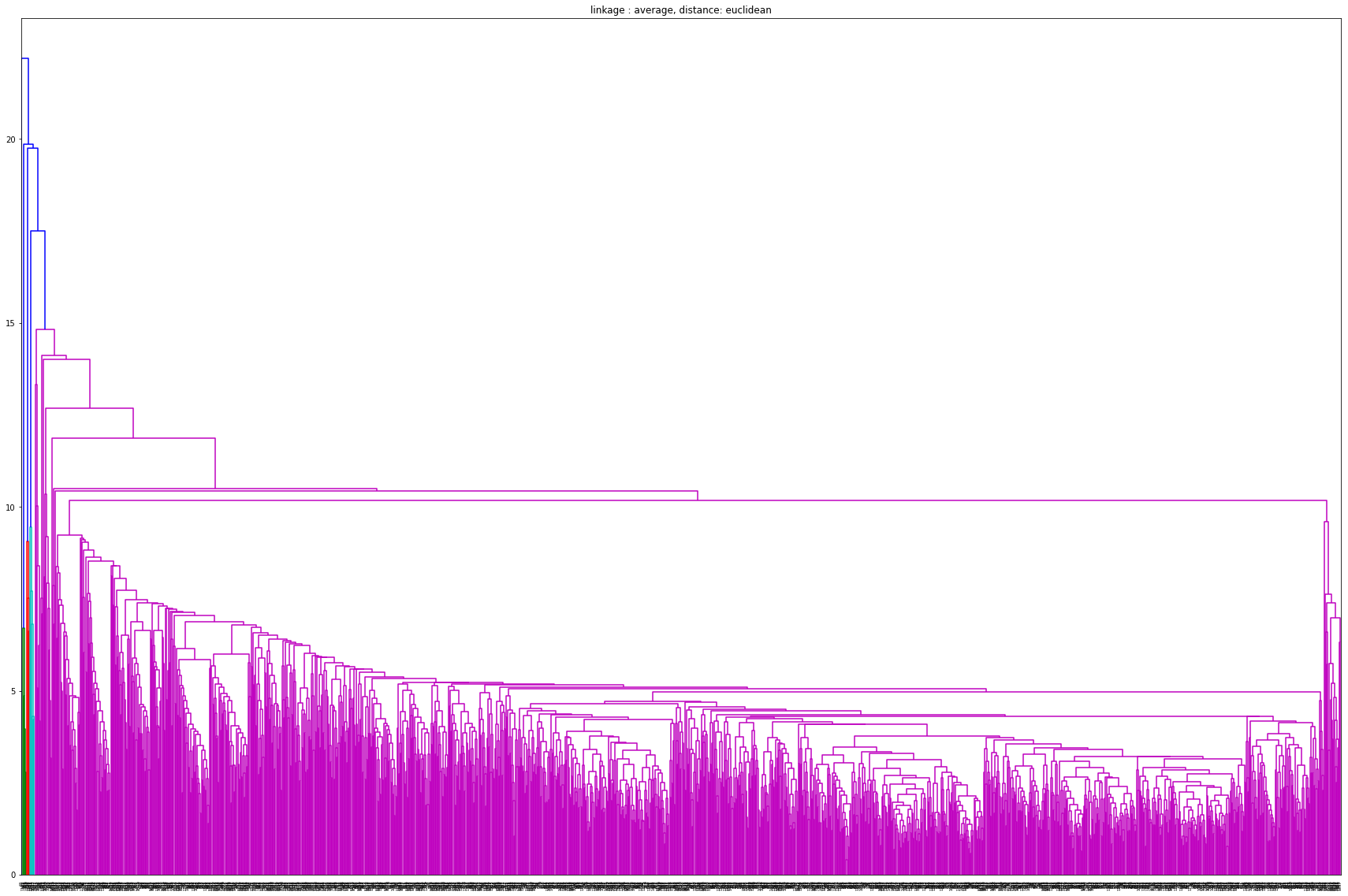
**Complete Linkage**

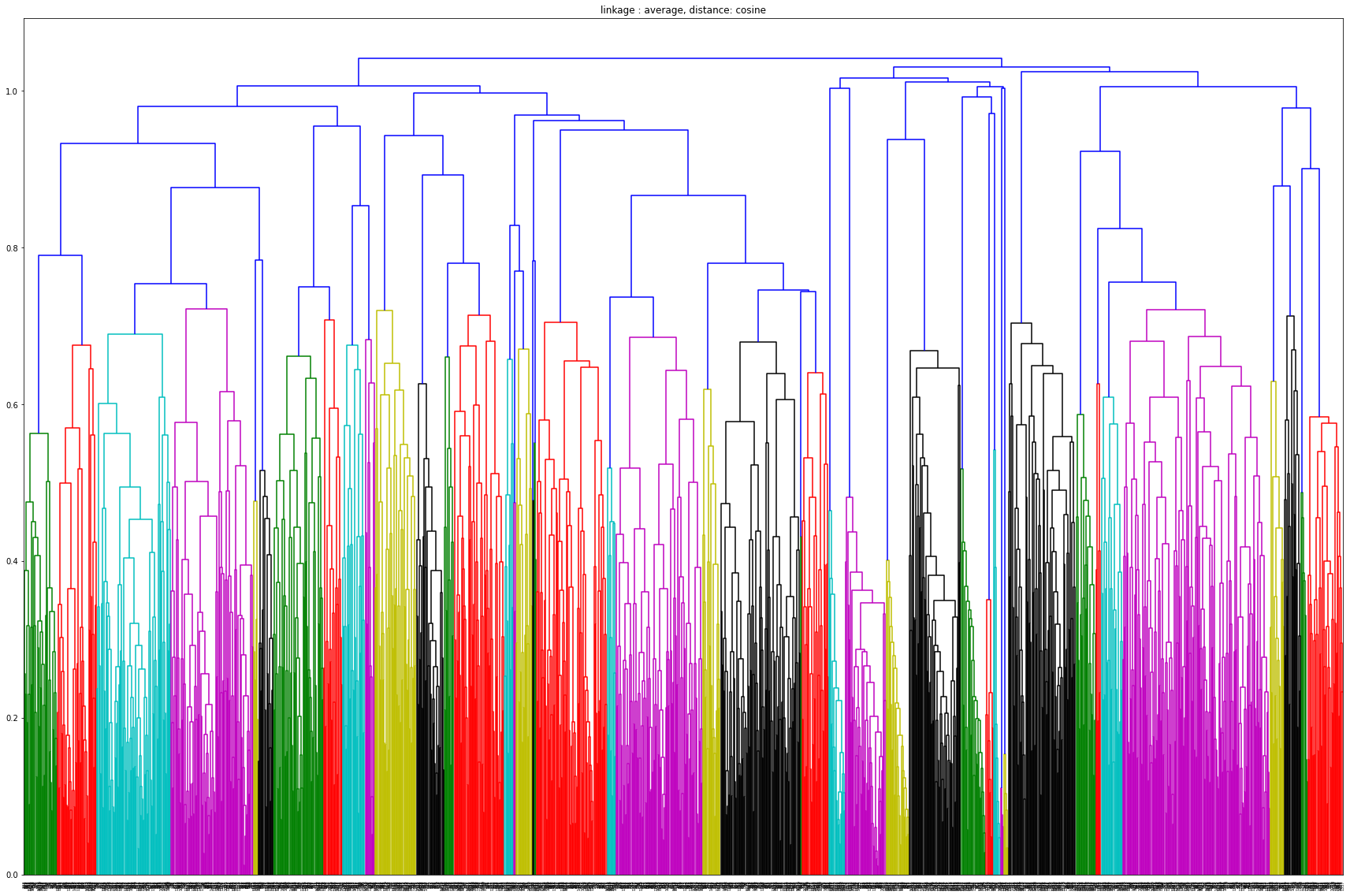
****

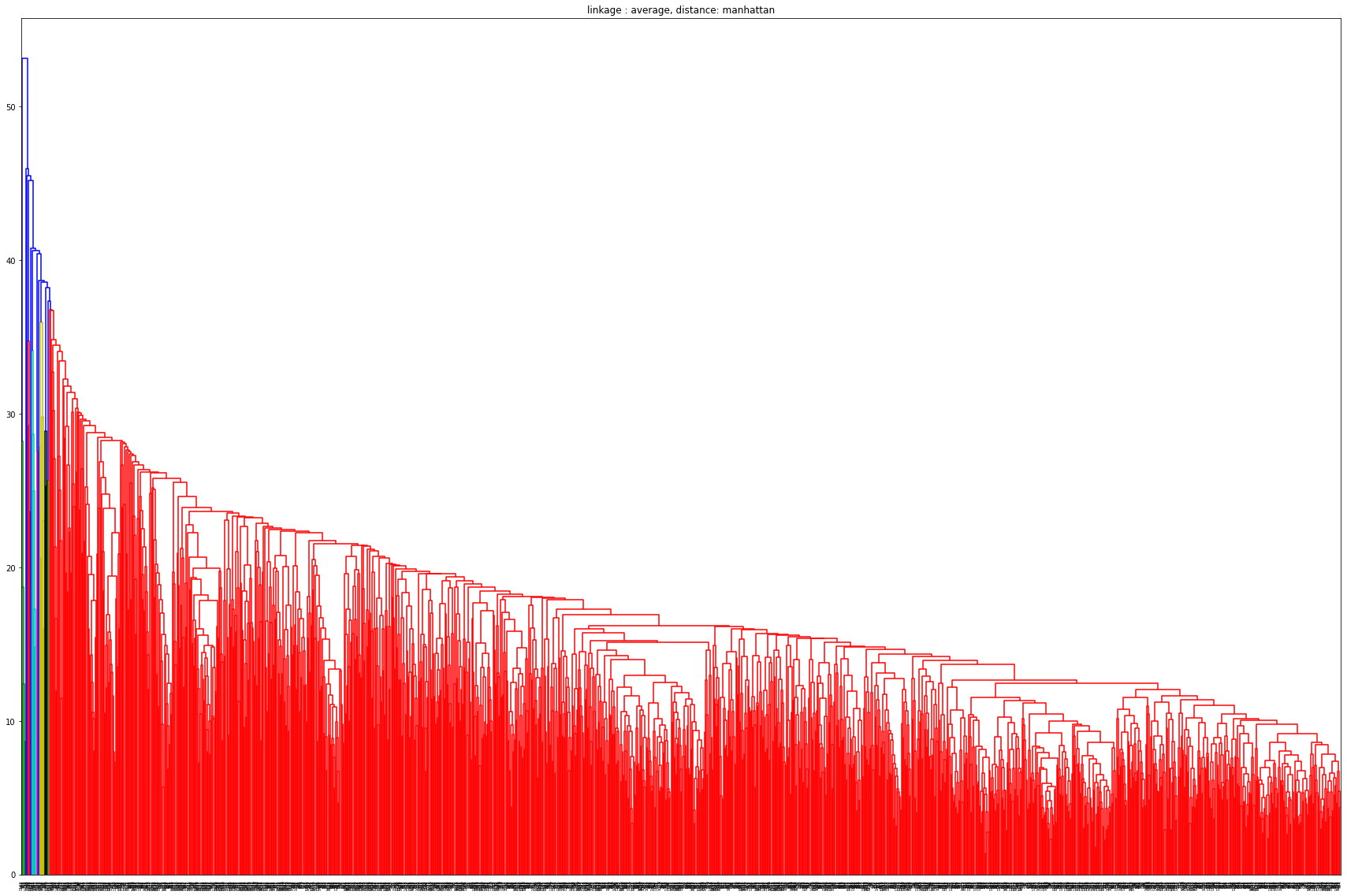
****

****

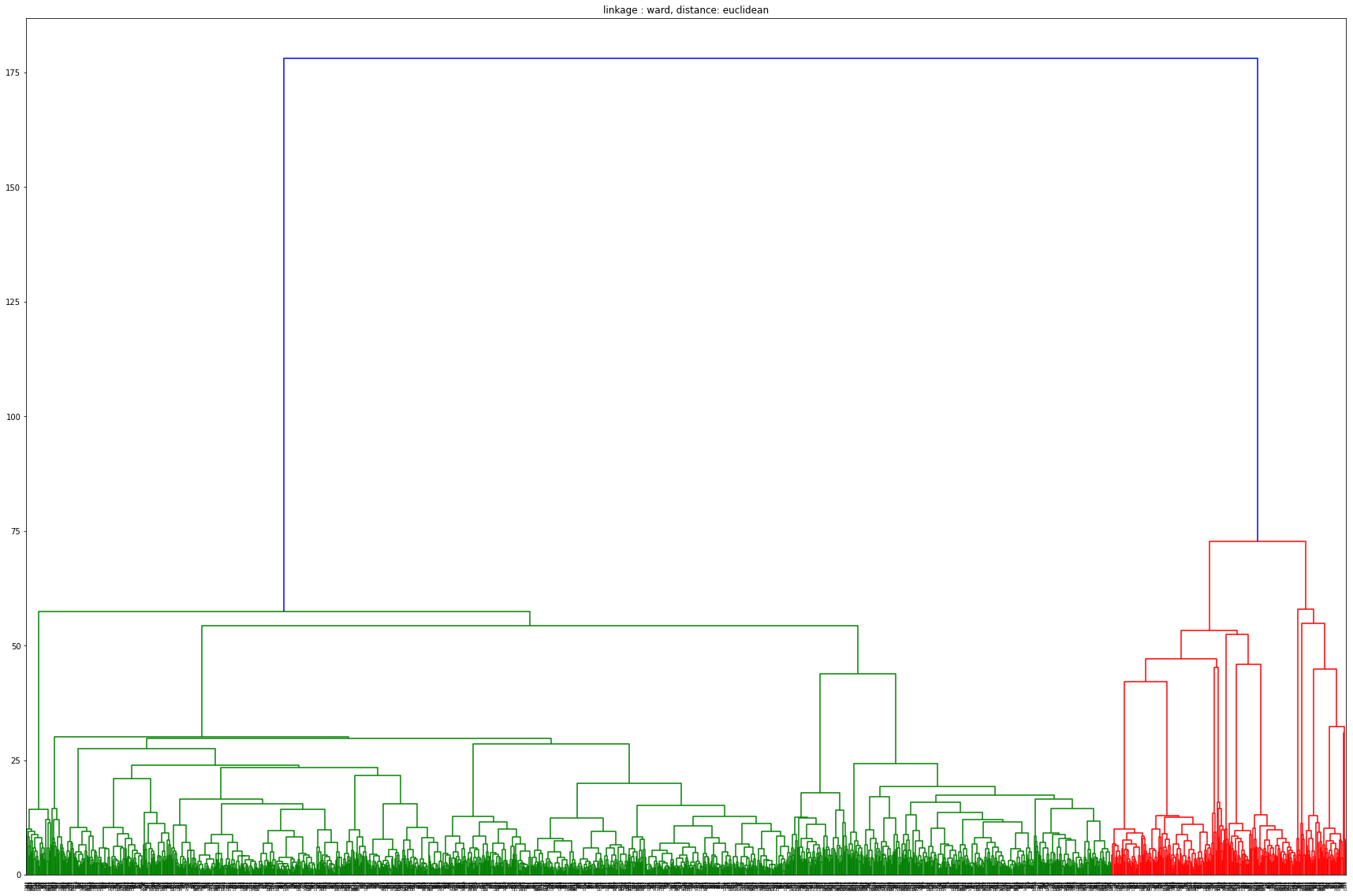
**Average Linkage**

****

****

****

Ward

****

Criteri di valutazione:

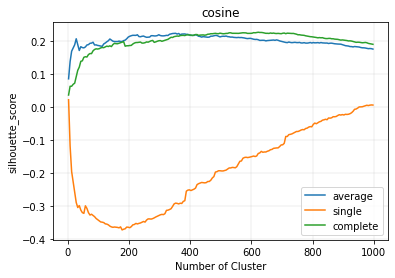
Due tipi di misure di convalida possono essere utilizzate per misurare le somiglianze tra le soluzioni di clustering: interne ed esterne. Il primo tipo di criterio misura gli attributi presi dai dati stessi e dai cluster formati, come la compattezza e la separabilità dei dati. Il secondo fa un confronto 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

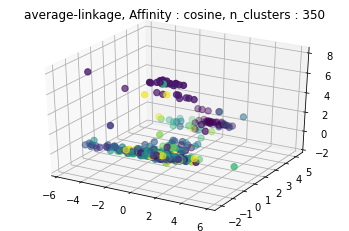
number of cluster of size 1 : 73

number of cluster of size 15 : 2

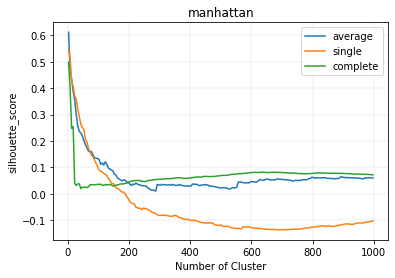
number of cluster of size 17 : 1

number of cluster of size 26 : 1

number of cluster of size 21 : 1



Manhattan

****

Build Model

MODEL OUTPUT

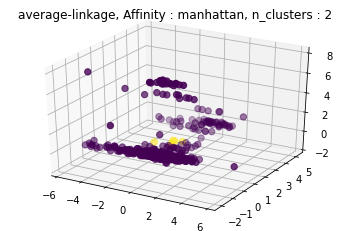
==================================================================

averageLinkage, Affinity : manhattan, n\_clusters : 2

Silhouette : 0.6114956890733572

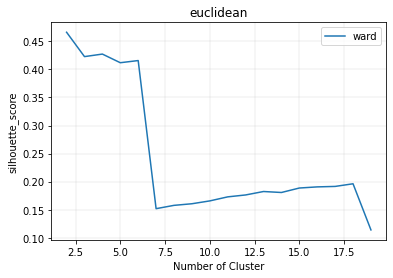
number of cluster of size 1565 : 1

number of cluster of size 5 : 1

****

|  |  |
| --- | --- |
|  |  |
| 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 |  |

Ward + Euclidean

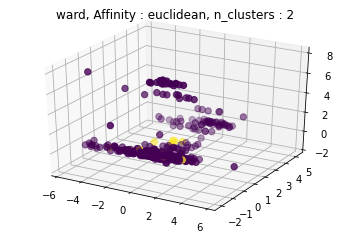
****

Build Model

Silhouette : 0.491189944795383

number of cluster of size 1522 : 1

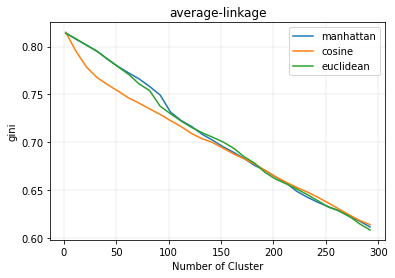
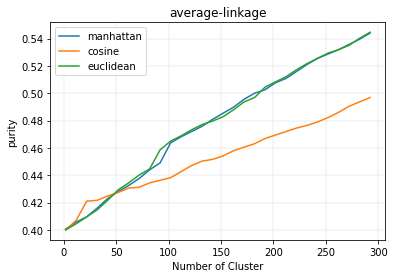
number of cluster of size 48 : 1

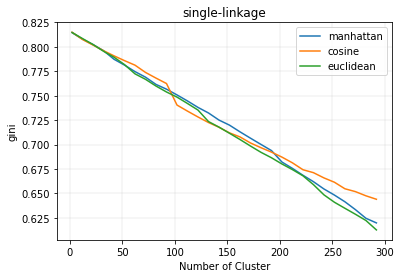
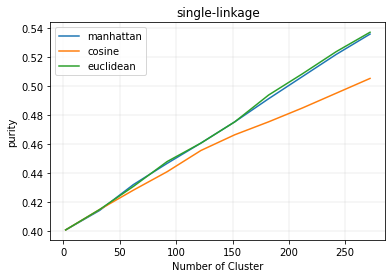
****

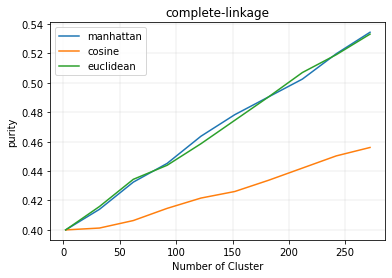
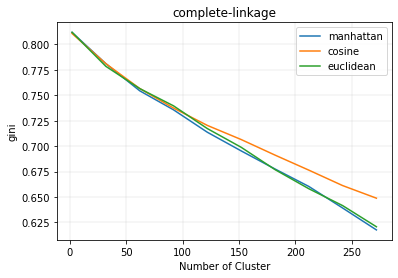
Criteri Esterni

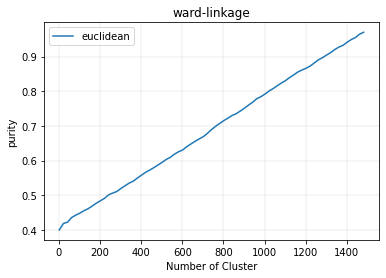
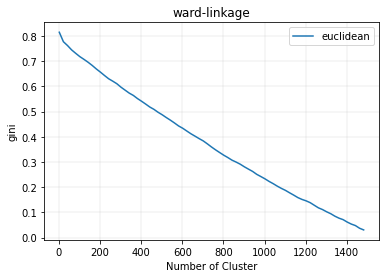
Per valutare i risultati del cluster scelto potrei utilizzare le variabili di output presenti nel dataset e un quindi un criterio di validazione esterna, tali criteri permettono di quantificare le somiglianze tra due soluzioni di clustering.

Gini e Purity









**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 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 loro

Soluzione: Indice OC

Con criteri esterni classici, è possibile solamente quantificare quanto sono simili due **soluzioni disgiunte**, dove ogni oggetto può essere etichettata 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.

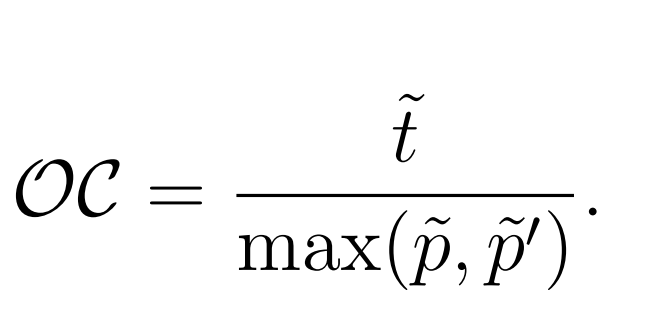
**Nota bene:**

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

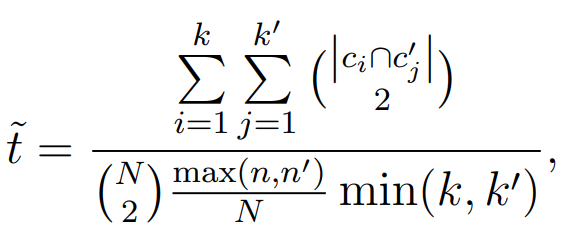
**Esempio:** il numero di cluster è 10 è il cluster predetto è etichettato con il numero  **7** allora in formato dummy sarà **.**

Un criterio utili allo scopo è l’indice di **Overlapped Cluster** (OC) [[1]](#footnote-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.

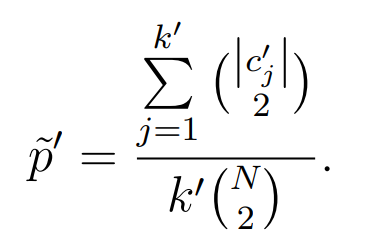
**Dettagli sul criterio OC:**



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

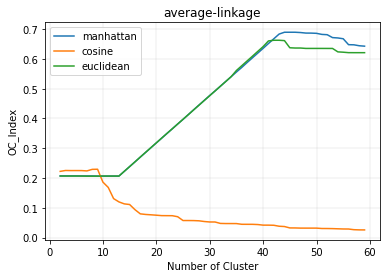
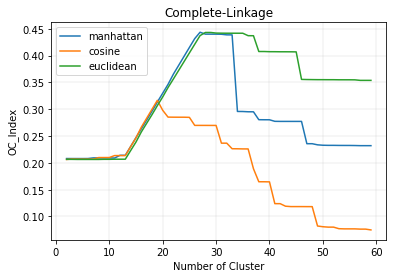


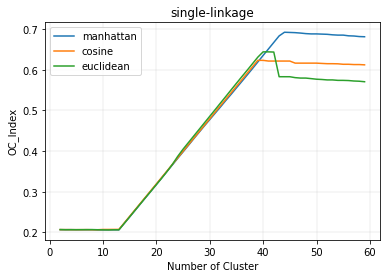
La probabilità di trovare una coppia di elementi in qualsiasi cluster per tutti i cluster viene stimata come

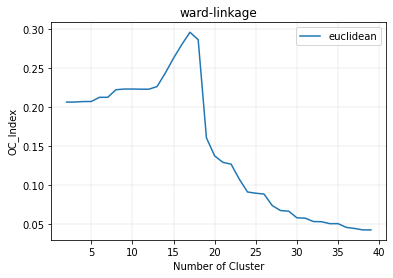


OC Index

Utilizzando come criterio esterno OC avremo i seguenti risultati.

****

****

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 0 mentre con la Mahnathan si arriva al “gomito” dopo i 40 cluster per poi descrescere.

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}

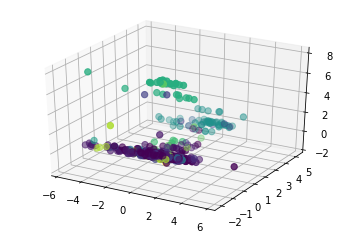
C\_12 size : 11 labels : {'FIBR\_PREDS': 1, 'JELUD\_TAH': 1, 'NO\_COMP': 1, 'OTEK\_LANC': 2, 'REC\_IM': 3, 'ZSN': 3}

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}

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}

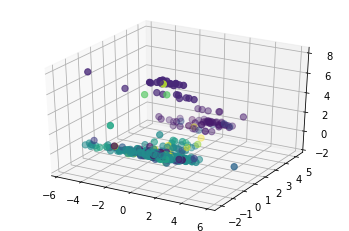
C\_23 size : 3 labels : {'LET\_IS': 1, 'RAZRIV': 1, 'REC\_IM': 1}

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}

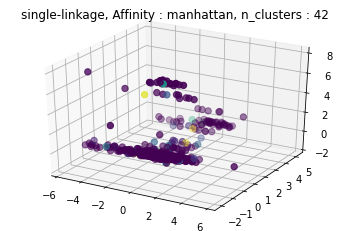
C\_37 size : 2 labels : {'A\_V\_BLOK': 1, 'ZSN': 1}

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}

****

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}

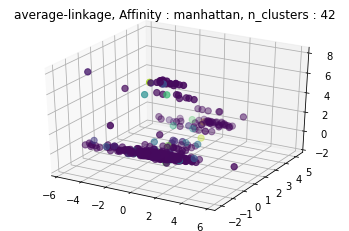
C\_37 size : 3 labels : {'FIBR\_PREDS': 1, 'REC\_IM': 1, 'ZSN': 1}

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}

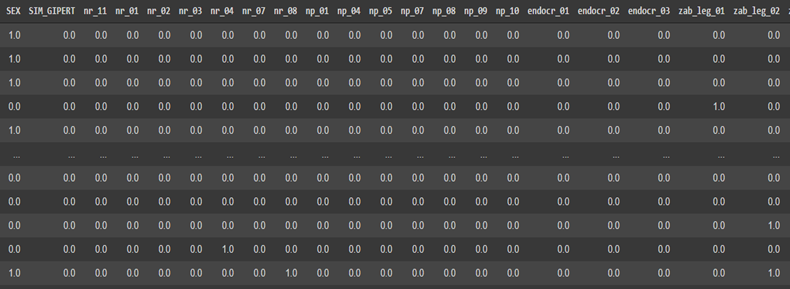
C\_41 size : 1 labels : {'NO\_COMP': 1}

****

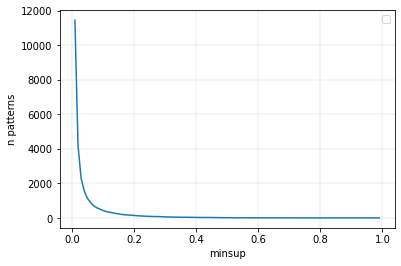
Estrazione di regole di Associazione

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

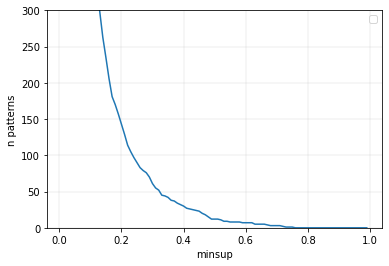
Il dataset ristretto alle soli variabili binari viene quindi considerato come un insieme di transazioni.

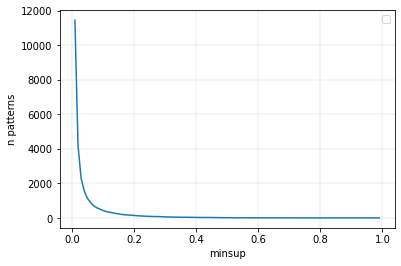


Eseguendo l’algoritmo Apriori per diversi valori di  **minsup** si ottiene il seguente grafico:



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 | ('ASP*S*n',) | 1182 |
| 1 | ('LID*S*n',) | 454 |
| 2 | ('GEPAR*S*n',) | 1123 |
| 3 | ('ritm*ecg*p\_07',) | 350 |
| 4 | ('SEX',) | 986 |
| 5 | ('ANT*CA*S\_n',) | 1060 |
| 6 | ('ritm*ecg*p\_01',) | 1052 |
| 7 | ('TRENT*S*n',) | 315 |
| 8 | ('ZSN',) | 364 |
| 9 | ('LET\_IS',) | 248 |
| 10 | ('ASP*S*n', 'GEPAR*S*n') | 923 |
| 11 | ('ASP*S*n', 'LID*S*n') | 355 |
| 12 | ('ASP*S*n', 'SEX') | 732 |
| 13 | ('ASP*S*n', 'ritm*ecg*p\_07') | 268 |
| 14 | ('GEPAR*S*n', 'LID*S*n') | 342 |
| 15 | ('GEPAR*S*n', 'SEX') | 716 |
| 16 | ('GEPAR*S*n', 'ritm*ecg*p\_07') | 246 |
| 17 | ('LID*S*n', 'SEX') | 321 |
| 18 | ('ANT*CA*S*n', 'ASP*S\_n') | 827 |
| 19 | ('ANT*CA*S*n', 'GEPAR*S\_n') | 758 |
| 20 | ('ANT*CA*S*n', 'LID*S\_n') | 301 |
| 21 | ('ANT*CA*S\_n', 'SEX') | 655 |
| 22 | ('ANT*CA*S*n', 'ritm*ecg*p*01') | 719 |
| 23 | ('ASP*S*n', 'ritm*ecg*p\_01') | 808 |
| 24 | ('GEPAR*S*n', 'ritm*ecg*p\_01') | 767 |
| 25 | ('LID*S*n', 'ritm*ecg*p\_01') | 312 |
| 26 | ('SEX', 'ritm*ecg*p\_01') | 700 |
| 27 | ('ANT*CA*S\_n', 'ZSN') | 258 |
| 28 | ('ASP*S*n', 'ZSN') | 276 |
| 29 | ('GEPAR*S*n', 'ZSN') | 249 |
| 30 | ('ASP*S*n', 'GEPAR*S*n', 'LID*S*n') | 291 |
| 31 | ('ASP*S*n', 'GEPAR*S*n', 'SEX') | 594 |
| 32 | ('ASP*S*n', 'LID*S*n', 'SEX') | 246 |
| 33 | ('GEPAR*S*n', 'LID*S*n', 'SEX') | 245 |
| 34 | ('ANT*CA*S*n', 'ASP*S*n', 'GEPAR*S\_n') | 627 |
| 35 | ('ANT*CA*S*n', 'ASP*S*n', 'LID*S\_n') | 240 |
| 36 | ('ANT*CA*S*n', 'ASP*S\_n', 'SEX') | 497 |
| 37 | ('ANT*CA*S*n', 'ASP*S*n', 'ritm*ecg*p*01') | 565 |
| 38 | ('ANT*CA*S*n', 'GEPAR*S\_n', 'SEX') | 472 |
| 39 | ('ANT*CA*S*n', 'GEPAR*S*n', 'ritm*ecg*p*01') | 517 |
| 40 | ('ANT*CA*S*n', 'SEX', 'ritm*ecg*p*01') | 472 |
| 41 | ('ASP*S*n', 'GEPAR*S*n', 'ritm*ecg*p\_01') | 640 |
| 42 | ('ASP*S*n', 'LID*S*n', 'ritm*ecg*p\_01') | 243 |
| 43 | ('ASP*S*n', 'SEX', 'ritm*ecg*p\_01') | 535 |
| 44 | ('GEPAR*S*n', 'LID*S*n', 'ritm*ecg*p\_01') | 244 |
| 45 | ('GEPAR*S*n', 'SEX', 'ritm*ecg*p\_01') | 517 |
| 46 | ('ANT*CA*S*n', 'ASP*S*n', 'GEPAR*S\_n', 'SEX') | 392 |
| 47 | ('ANT*CA*S*n', 'ASP*S*n', 'GEPAR*S*n', 'ritm*ecg*p*01') | 433 |
| 48 | ('ANT*CA*S*n', 'ASP*S*n', 'SEX', 'ritm*ecg*p*01') | 365 |
| 49 | ('ANT*CA*S*n', 'GEPAR*S*n', 'SEX', 'ritm*ecg*p*01') | 344 |
| 50 | ('ASP*S*n', 'GEPAR*S*n', 'SEX', 'ritm*ecg*p\_01') | 438 |
| 51 | ('ANT*CA*S*n', 'ASP*S*n', 'GEPAR*S*n', 'SEX', 'ritm*ecg*p*01') | 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},

{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 manahattan
* 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 cluste**r 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 dendogramma 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 i**n modo da bilanciare il dataset sarà possibile ottenere risultati più interessanti con la cluster analsysis

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

**{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

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 [↑](#footnote-ref-1)