Analisi dei dati per la Sicurezza

Gabriele Patta

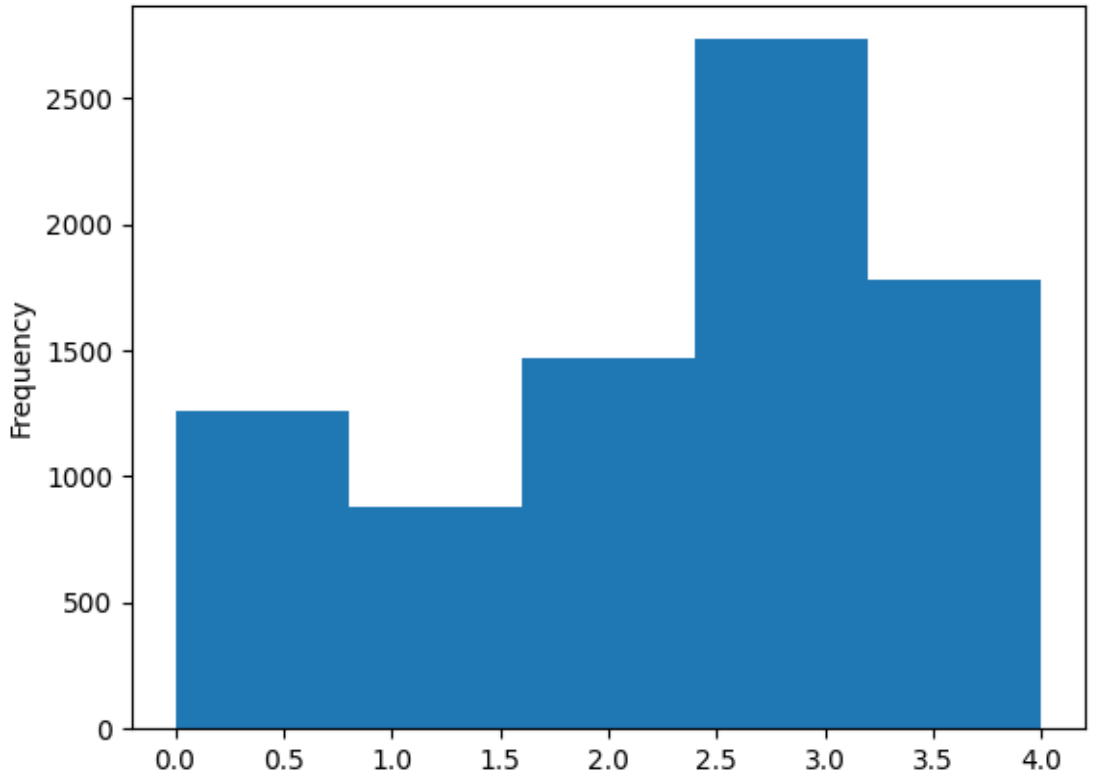
Matricola: 756410

AA. 2022/2023

Dataset: CICMalDroid 2020 del *Canadian Institute for Cybersecurity (CIC)*

- Il *dataset* utilizzato è stato realizzato nel 2020 e contiene 11598 campioni Android collezionati da svariate fonti come VirusTotal, Contagio *security blog*, AMD ed altri *dataset*.
- Il *dataset* si compone dei seguenti elementi:
  - 8118 esempi
  - 41 attributi (40 attributi numerici che fungono da variabili indipendenti ed 1 rappresenta la classe, nonchè variabile dipendente)
- Nel *dataset* è possibile apprezzare 4 differenti esempi di minacce; N.B: "*Benign*" identifica la categoria nella quale sono contenute le applicazioni non elencate nelle tipologie precedenti.
  - *Adware* (1.256 esempi)
  - *Banking* (877 esempi)
  - SMS *malware* (1.470 esempi)
  - *Riskware* (2.733 esempi)
  - *Benign* (1.782 esempi)

Distribuzione dei dati (*train*)



Inizialmente si è inteso valutare la distribuzione dei dati delle classi a cui appartengono gli esempi del *dataset* utilizzato.
Tale distribuzione risulta abbastanza bilanciata.

Feature selection

- Successivamente si è provveduto ad individuare un sottoinsieme di *feature* maggiormente significative per poter svolgere la classificazione.
  - Le *feature* sono state individuate attraverso la metrica di *Mutual information*:
    - 1 – fs_access(write)
    - 2 – mmap2
    - 3 – network_access
    - 4 – create_folder
    - 5 – unlink
    - 6 – device_access
    - 7 – rename
    - 8 – munmap
    - 9 – mkdir
    - 10 – fs_pipe_access

Feature selection

fs_access(write)	mmap2	network_access	create_folder	unlink	device_access	rename	munmap	mkdir	fs_pipe_access	Class
6	72	0	3	0	2	0	42	3	2	3
6	58	0	3	3	2	3	28	3	1	3
6	83	0	3	0	2	0	27	3	1	3
8	393	11	3	8	12	6	175	5	8	1
0	0	0	0	0	0	0	0	0	0	0
6	32	0	2	0	2	0	2	2	2	2
0	0	0	0	0	0	0	0	0	0	0
6	83	0	3	0	2	0	27	3	1	3
6	114	0	3	0	2	0	40	3	1	1
6	42	0	3	0	2	0	6	3	1	2
6	244	0	2	0	2	0	180	2	1	2
0	0	0	0	0	0	0	0	0	0	0
6	36	0	3	0	2	0	3	3	2	4
0	0	0	0	0	0	0	0	0	0	0
6	351	10	5	2	34	1	186	5	1	3

Top 10 *features*

mmap2	access	pread64	open	munmap	stat64	gettid	fcntl64	fs_access	mprotect	Class
0,004158	0,000382	0,000701	0,000396	0,002472	0,001057	0,001018	0,000122	0,000379	0,001182	3
0,003350	0,000763	0,000443	0,000443	0,001648	0,001321	0,001018	0,000102	0,000379	0,000985	3
0,004793	0,000636	0,000295	0,000602	0,001589	0,001057	0,001239	0,000163	0,001010	0,002812	3
0,022696	0,011318	0,000941	0,014204	0,010298	0,005973	0,013411	0,006107	0,007452	0,008419	1
0,000000	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000	0
0,001848	0,000191	0,000018	0,000364	0,000118	0,000581	0,000929	0,000102	0,000379	0,000609	2
0,000000	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000	0
0,004793	0,000636	0,000295	0,000602	0,001589	0,001057	0,001239	0,000163	0,001010	0,002741	3
0,006584	0,000699	0,000332	0,000934	0,002354	0,001427	0,001416	0,000204	0,002653	0,003386	1
0,002426	0,000382	0,000018	0,000934	0,000353	0,001744	0,001239	0,000835	0,001263	0,002687	2
0,014091	0,000318	0,002951	0,000649	0,010593	0,001850	0,001151	0,000305	0,001895	0,006323	2
0,000000	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000	0
0,002079	0,000509	0,000018	0,000412	0,000177	0,000951	0,000929	0,000102	0,000505	0,001039	4
0,000000	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000	0
0,020270	0,001399	0,000535	0,019667	0,010946	0,002907	0,013897	0,001852	0,006442	0,008545	3

Top 10 *features* con MinMaxScaler

Dapprima è stato eseguito lo *scaling* dei dati attraverso il MinMaxScaler. In seguito è stato svolto il *ranking* delle *feature* e ne sono state identificate un *set* differente:

mmap2 - access – pread64 – open – munmap – stat64 – gettid – fcntl64 – fs_access - mprotect

PCA (*Principal Component Analysis*)

```
Listing 10 top important principal components from given data frame:
          pc_1        pc_2        pc_3        pc_4        pc_5     \
0    -687.337208 -496.707003  109.701079 -146.449233 -139.467978
1    -682.666562 -497.602110  103.979975 -155.888021 -143.528650
2    -685.371298 -460.248059  119.529289 -136.407983  -72.652283
3    -601.347344  673.877187  372.305740  135.954684  301.428264
4    -693.863056 -550.544978   99.461397 -182.553066 -173.926583
...          ...         ...         ...         ...         ...
8113 -626.344143  505.136887  480.084303   48.872541  404.971729
8114 -681.117008 -458.827284  124.807785 -130.967669  -66.727776
8115 -681.687209 -433.759494  133.934033 -120.378274  -13.258364
8116 -684.565669 -417.173707   66.372137 -112.601509  -72.604724
8117 -678.402114 -370.613237  113.487592  -51.143807  -67.645219

          pc_6       pc_7        pc_8       pc_9       pc_10  Class
0     125.072586  62.908703  -41.193217  16.841220  -25.414781      3
1     141.804145  56.797013  -48.273274  18.741779  -36.968284      3
2     81.101334   54.048140  -65.722038  19.117916  -46.714294      3
3     276.303515 -13.756801  292.402686 -46.842235  114.650757      1
4     200.889161  64.169247  -74.787671  31.713748  -64.398370      0
...          ...         ...         ...         ...         ...     ...
8113  431.723674  95.679189  330.829788 -59.237662  101.258840      3
8114   70.848567  51.861899  -68.443235  19.071909  -43.899802      3
8115   17.928817  63.562970  -88.292361  22.293136  -64.094719      3
8116   64.694088  27.055023  -62.375213  13.626419  -63.814402      2
8116   29.345591  16.955287  -95.374973  21.900799  -65.268589      4

[8118 rows x 11 columns]
```

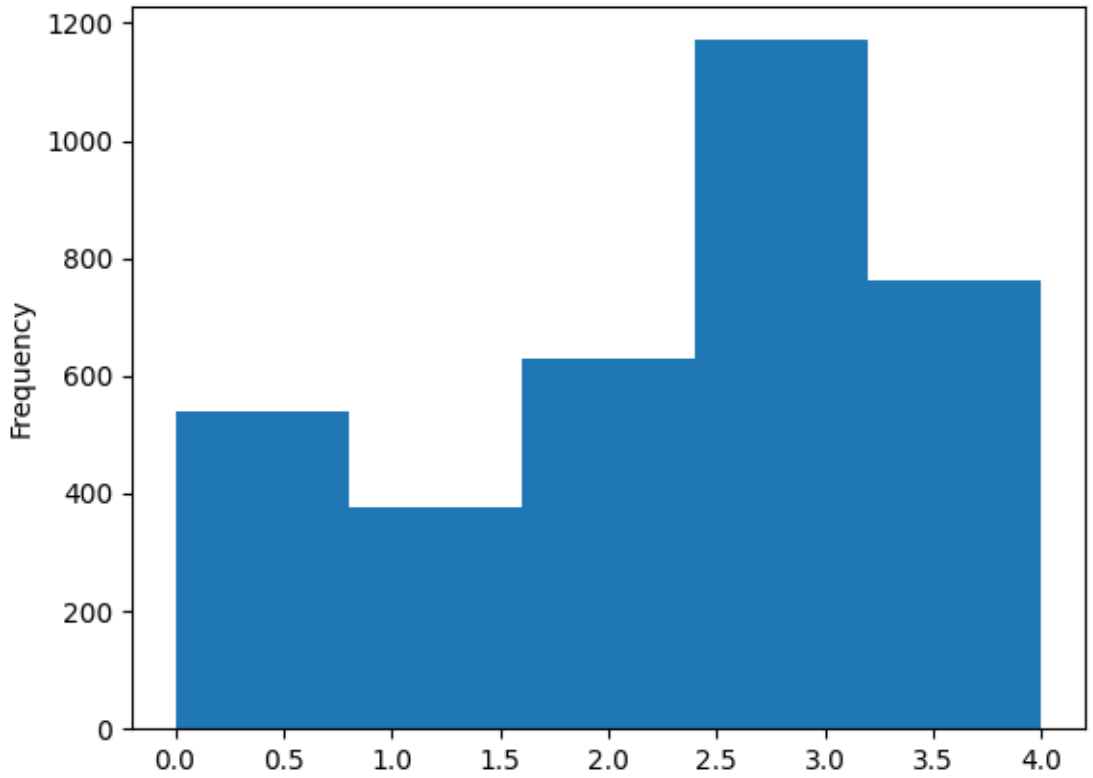
- È stata applicata la tecnica denominata *Principal Component Analysis* (PCA) per eseguire la trasformazione degli attributi in nostro possesso in componenti principali.

Stratified K-Fold Cross Validation

- Si è provveduto ad individuare la migliore configurazione di parametri per l'algoritmo di *Data mining* che è stato utilizzato, l'algoritmo C4.5

- Feature Ranking by MI: Best criterion = gini - Best N = 30 and Best CV F1 score = 0.7971574251705963
- Feature Ranking by PCA: Best criterion = gini - Best N = 20 and Best CV F1 score = 0.7532215794439903

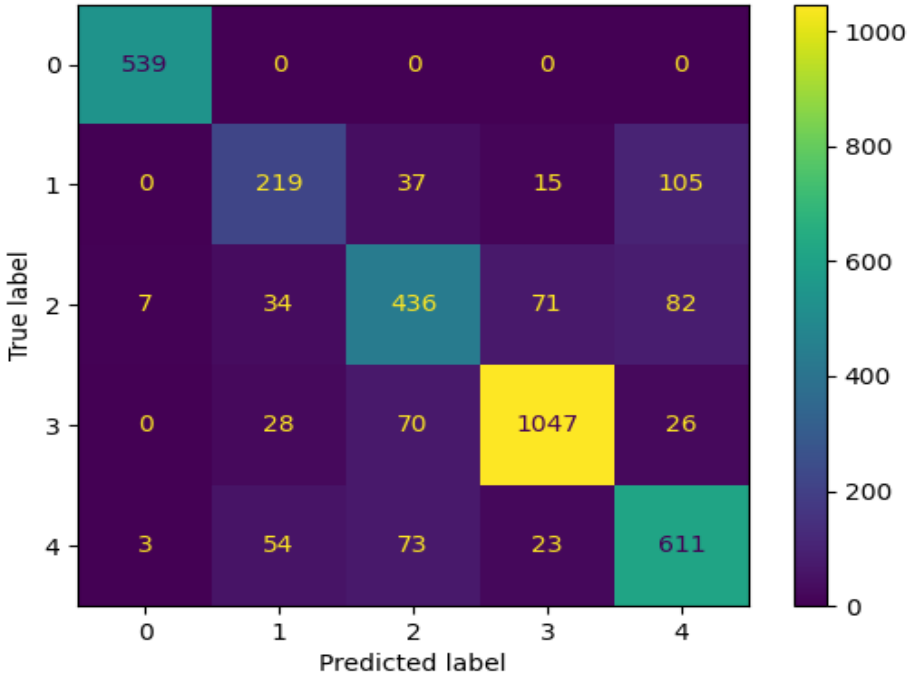
Distribuzione dei dati (*test*)



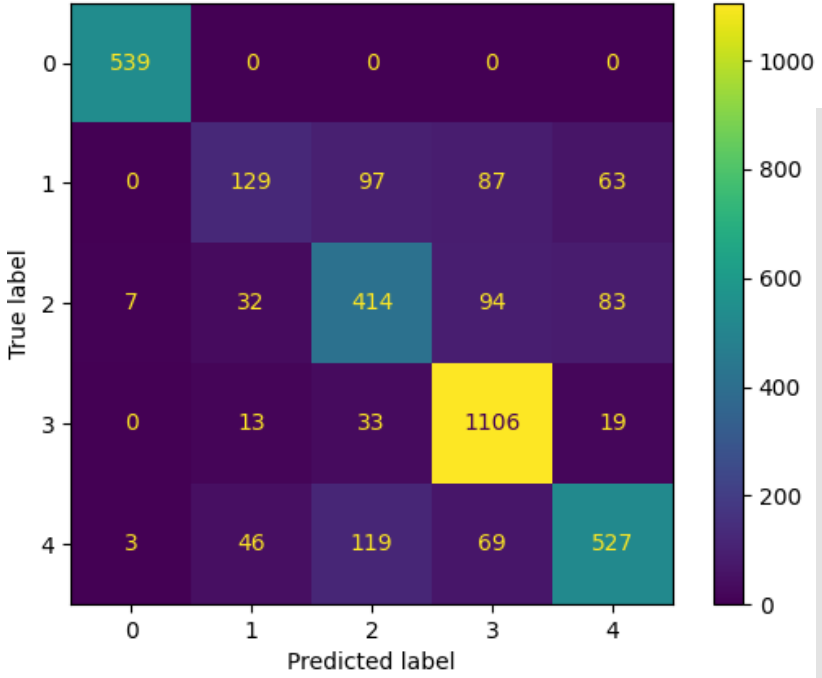
Il *dataset* di *test* presenta la stessa distribuzione dei dati del *training set* analizzata in precedenza. Sono presenti 3480 esempi, inoltre sono presenti 40 attributi + 1 per la classe.

Matrice di confusione e report di classificazione

 # Mutual info (sinistra)
PCA (destra)



	precision	recall	f1-score	support
0	0.98	1.00	0.99	539
1	0.65	0.58	0.62	376
2	0.71	0.69	0.70	630
3	0.91	0.89	0.90	1171
4	0.74	0.80	0.77	764
accuracy			0.82	3480
macro avg	0.80	0.79	0.80	3480
weighted avg	0.82	0.82	0.82	3480



	precision	recall	f1-score	support
0	0.98	1.00	0.99	539
1	0.59	0.34	0.43	376
2	0.62	0.66	0.64	630
3	0.82	0.94	0.88	1171
4	0.76	0.69	0.72	764
accuracy			0.78	3480
macro avg	0.75	0.73	0.73	3480
weighted avg	0.77	0.78	0.77	3480

I risultati migliori sono stati ottenuti con la configurazione *Mutual information*. Per quanto riguarda invece l'utilizzo della PCA, questa tecnica ha apportato dei risultati leggermente inferiori.

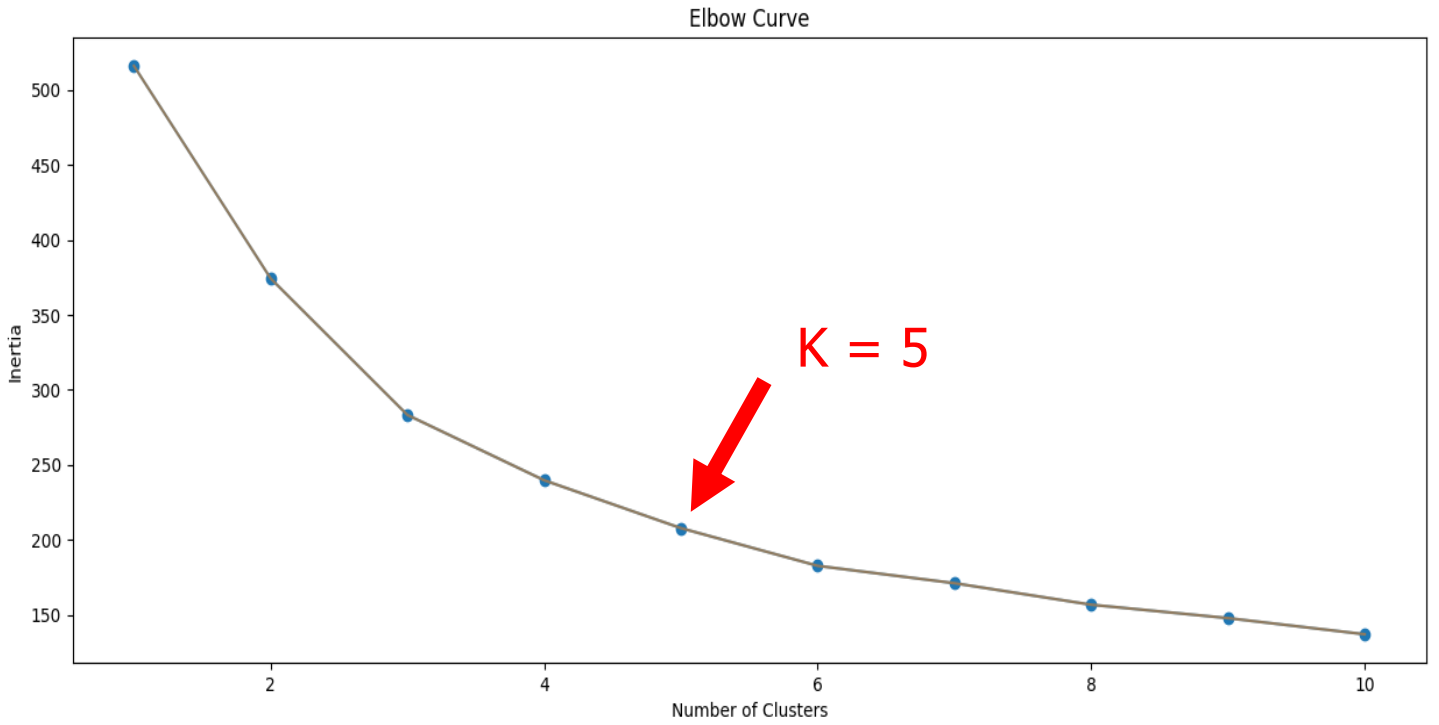
Clustering K-Means (non supervisionato)

Nell'algoritmo di *clustering* K-Means vengono utilizzati due parametri di particolare rilevanza:

- n_init: Determina il numero di volte in cui l'algoritmo viene eseguito con differenti centroidi iniziali (nel caso in esame tale valore sarà fissato a 10).
- inertia: Rappresenta la somma delle distanze quadratiche dei campioni dai loro centroidi più vicini; valori più bassi indicano assegnazioni dei *cluster* migliori.

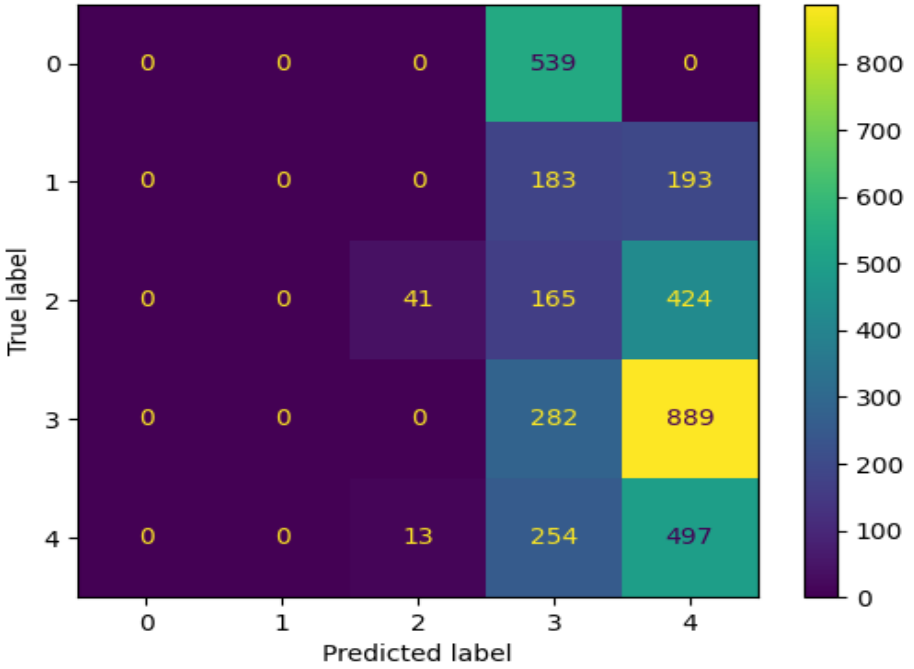
Al fine di ottimizzare l'esecuzione dell'algoritmo si è provveduto ad eseguire in un primo momento lo *scaling* dei dati attraverso l'algoritmo MinMaxScaler. Questo ha permesso all'algoritmo K-Means di poter operare con un *range* limitato in cui ricadono i dati.

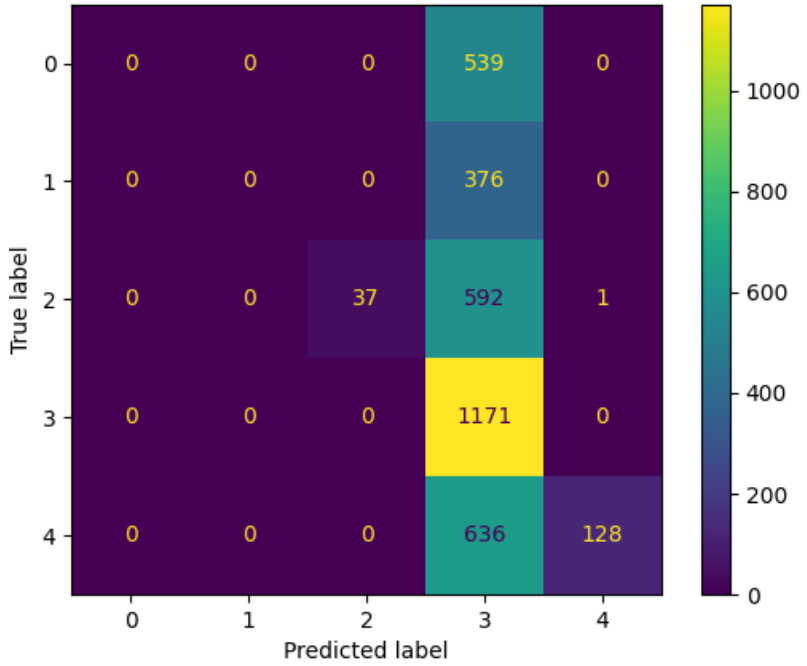
Valutazione dell'*Elbow curve*



Allo scopo di scegliere il numero ottimale di *cluster*, è possibile identificare sul grafico, il punto di "gomito" nella curva. Esso rappresenta il punto limite dove l'aggiunta di ulteriori *cluster* non diminuisce significativamente l'*inertia*. Inoltre attraverso questo punto è possibile ottenere un buon compromesso tra la complessità del modello (numero di *cluster*) e le prestazioni (*inertia*).

Matrice di confusione e report di classificazione






```
Optimal number of clusters: 5
Inertia: 207.90445087796226

Classification report:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00       539
           1       0.00      0.00      0.00       376
           2       0.76      0.07      0.12       630
           3       0.20      0.24      0.22      1171
           4       0.25      0.65      0.36       764

    accuracy                           0.24      3480
   macro avg       0.24      0.19      0.14      3480
weighted avg       0.26      0.24      0.17      3480
```

```
Optimal number of clusters: 5
Inertia: 207.90445087796226

Classification report:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00       539
           1       0.00      0.00      0.00       376
           2       1.00      0.06      0.11       630
           3       0.35      1.00      0.52      1171
           4       0.99      0.17      0.29       764

    accuracy                           0.38      3480
   macro avg       0.47      0.25      0.18      3480
weighted avg       0.52      0.38      0.26      3480
```

In questo caso è stato utilizzato l'algoritmo NearCentroid() per l'addestramento del modello.

Successivamente è stato impiegato l'algoritmo Kmeans() per l'addestramento del modello.

Conclusioni

- In questo progetto sono stati impiegati due algoritmi differenti allo scopo di addestrare il modello di *clustering.* In entrambi i casi il modello di classificazione ha ottenuto delle prestazioni non molto elevate, con valori di precisione, *recall* e F1-*score* molto bassi per la maggior parte delle classi. Tale eventualità indica che il modello non è in grado di fare delle previsioni accurate per la maggior parte delle istanze contenute nel *dataset* di *test.*

- Sono presenti un elevato numero di falsi positivi ed inoltre l'*accuracy* individuata in entrambi i casi (ovvero la proporzione delle predizioni corrette rispetto al totale delle predizioni) non supera lo 0,4.

Grazie per l'attenzione!