0.1 Soluzione

Per rispondere al problema si è scelto di utilizzare il tool JMP per l'analisi del dataset al fine di ridurne il numero di dati. Dopo un'analisi preliminare sui dati, la riduzione effettiva del dataset avviene utilizzando due tecniche: Principal Component Analysis (PCA) e clustering.

0.1.1 Analisi preliminare

L'analisi preliminare prevede, dopo aver importato il file nel tool, di analizzare le feature presenti, al fine di eliminare quelle prive di contenuto informativo. A tale scopo si è proceduto all'analisi delle distribuzioni di tali attributi osservando, in particolare, il coefficiente di variazione (CV) di ognuna. Il coefficiente di variazione è la normalizzazione della varianza con la media. Se il coefficiente di variazione è nullo, il parametro misurato è costante, e quindi si sceglie di non includere quella feature nelle successive analisi.

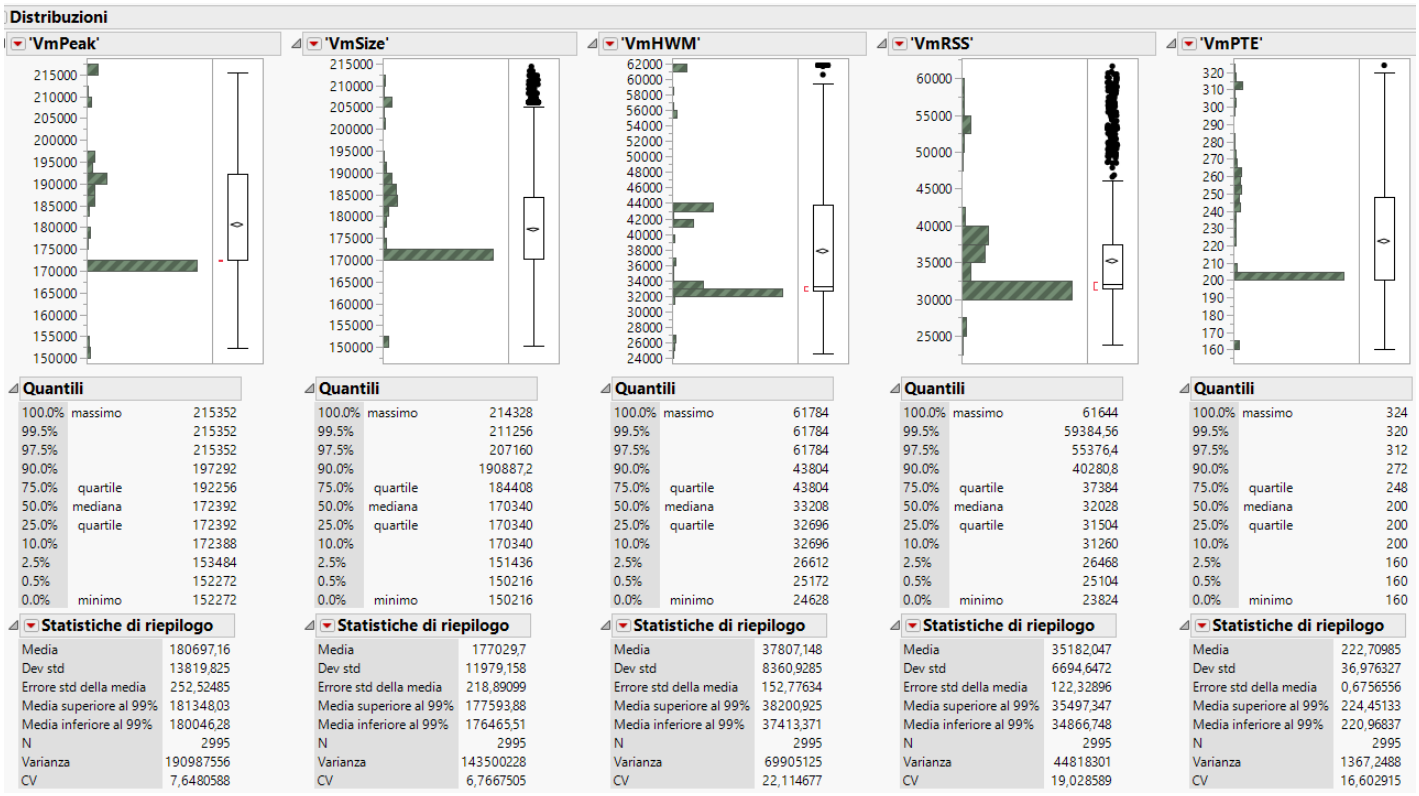

Figura 1: Distribuzioni delle feature

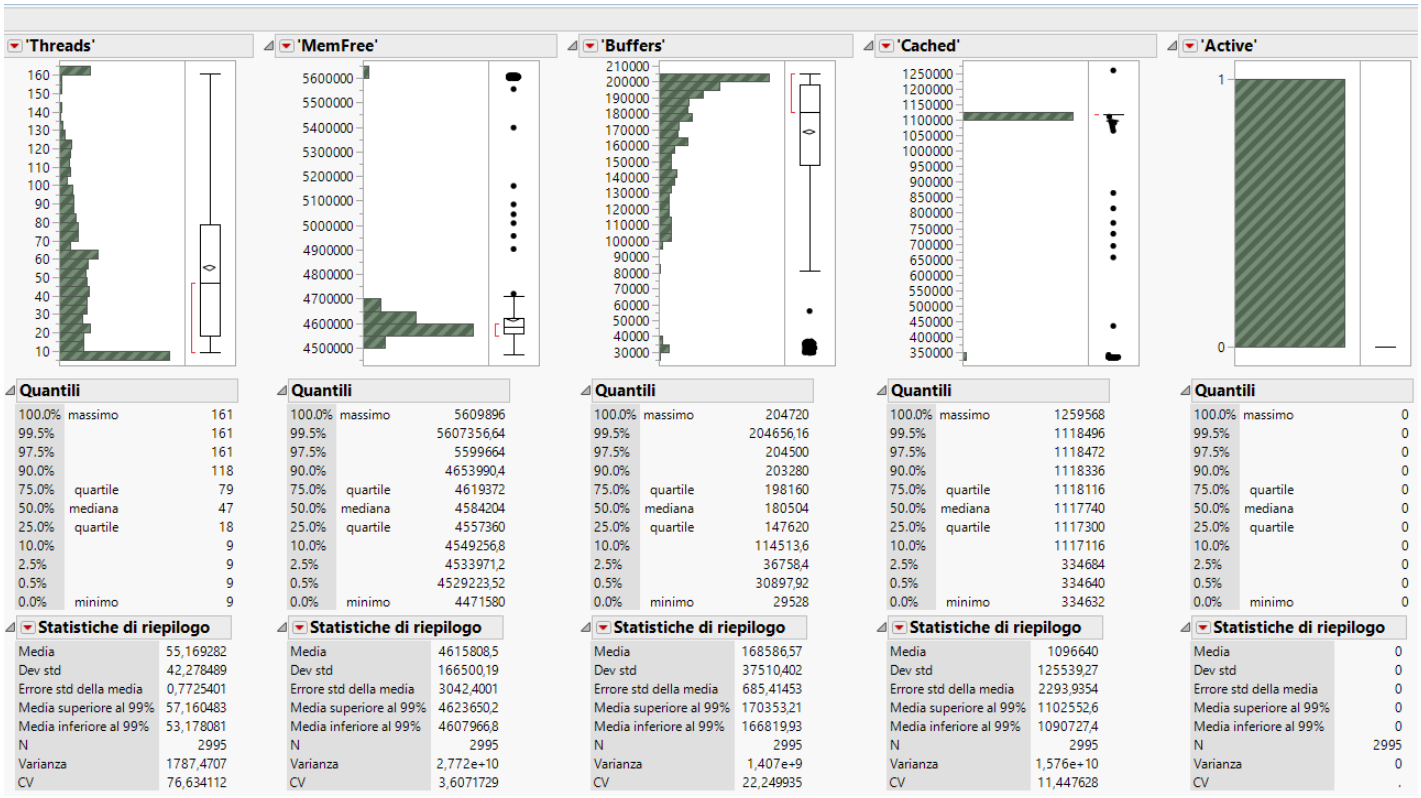

Figura 2: Distribuzioni delle feature

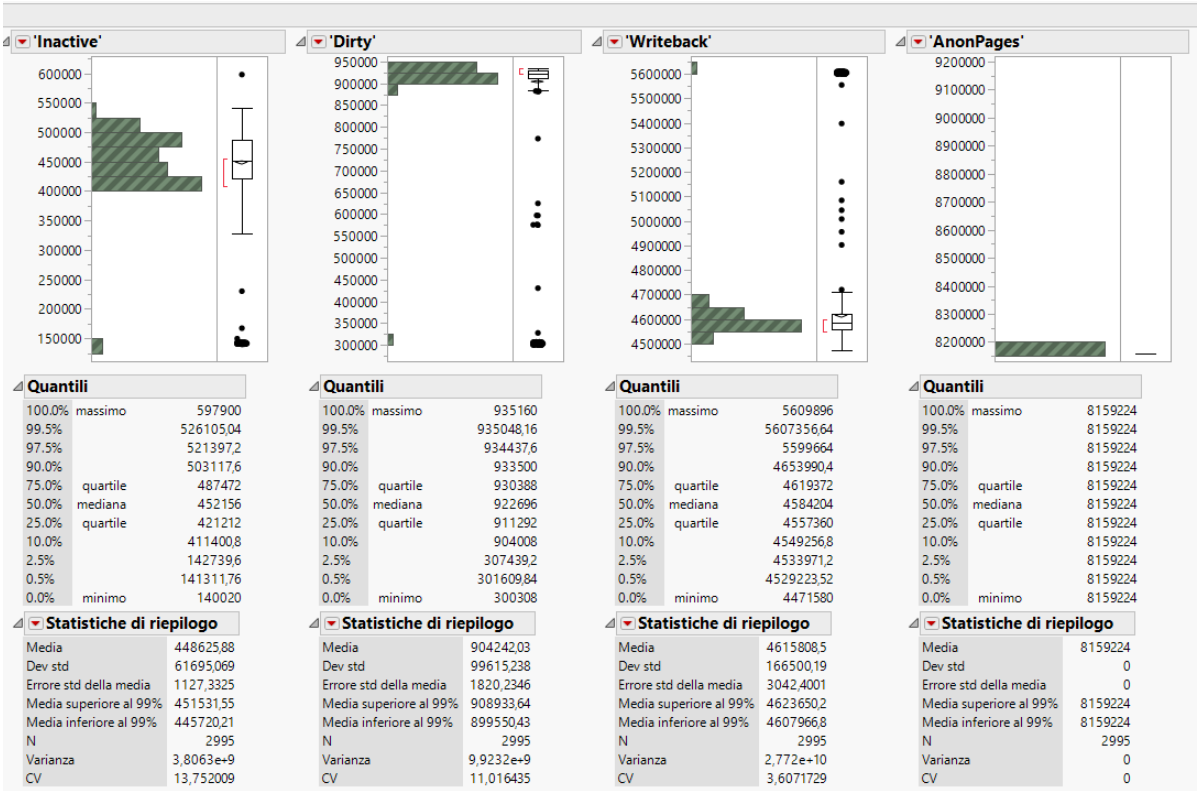

Figura 3: Distribuzioni delle feature

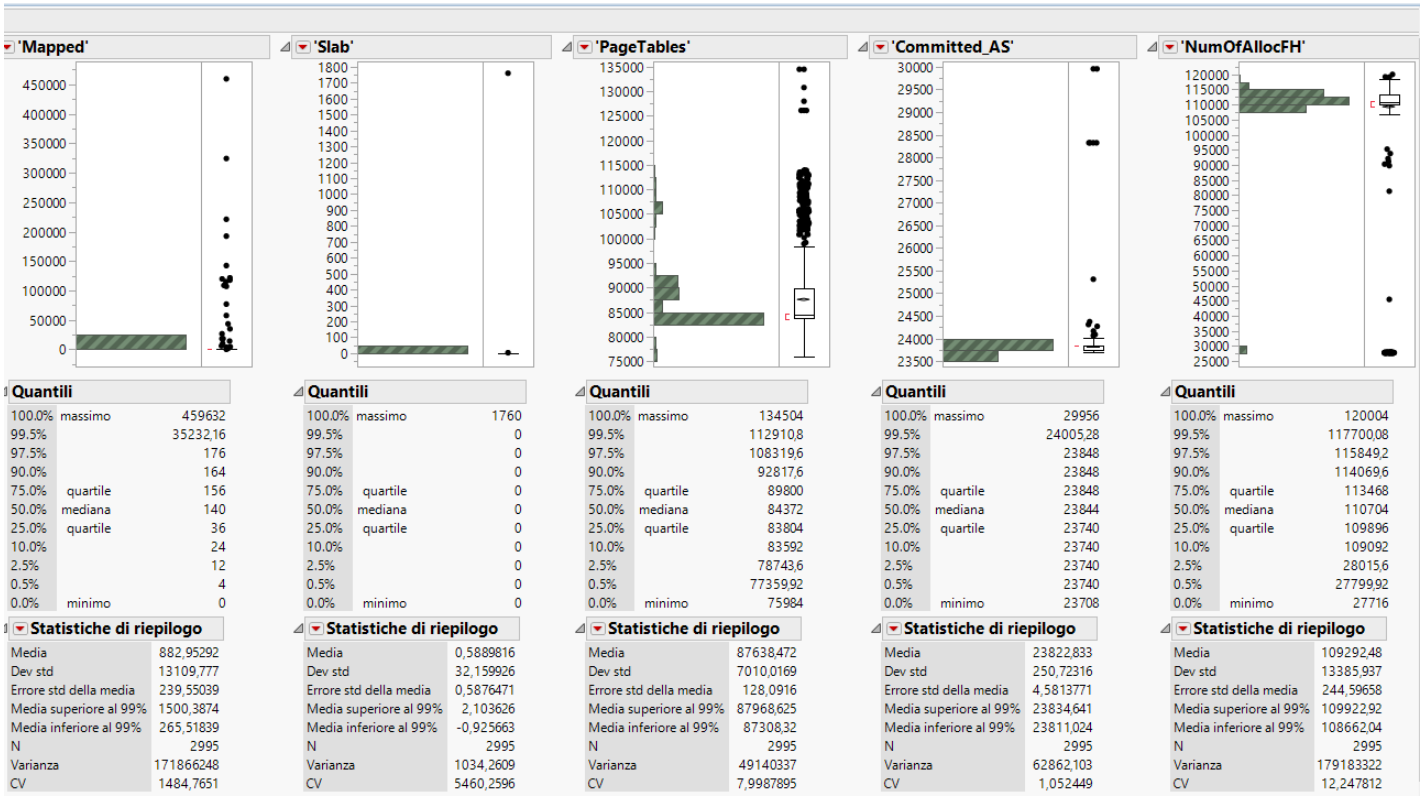

Figura 4: Distribuzioni delle feature

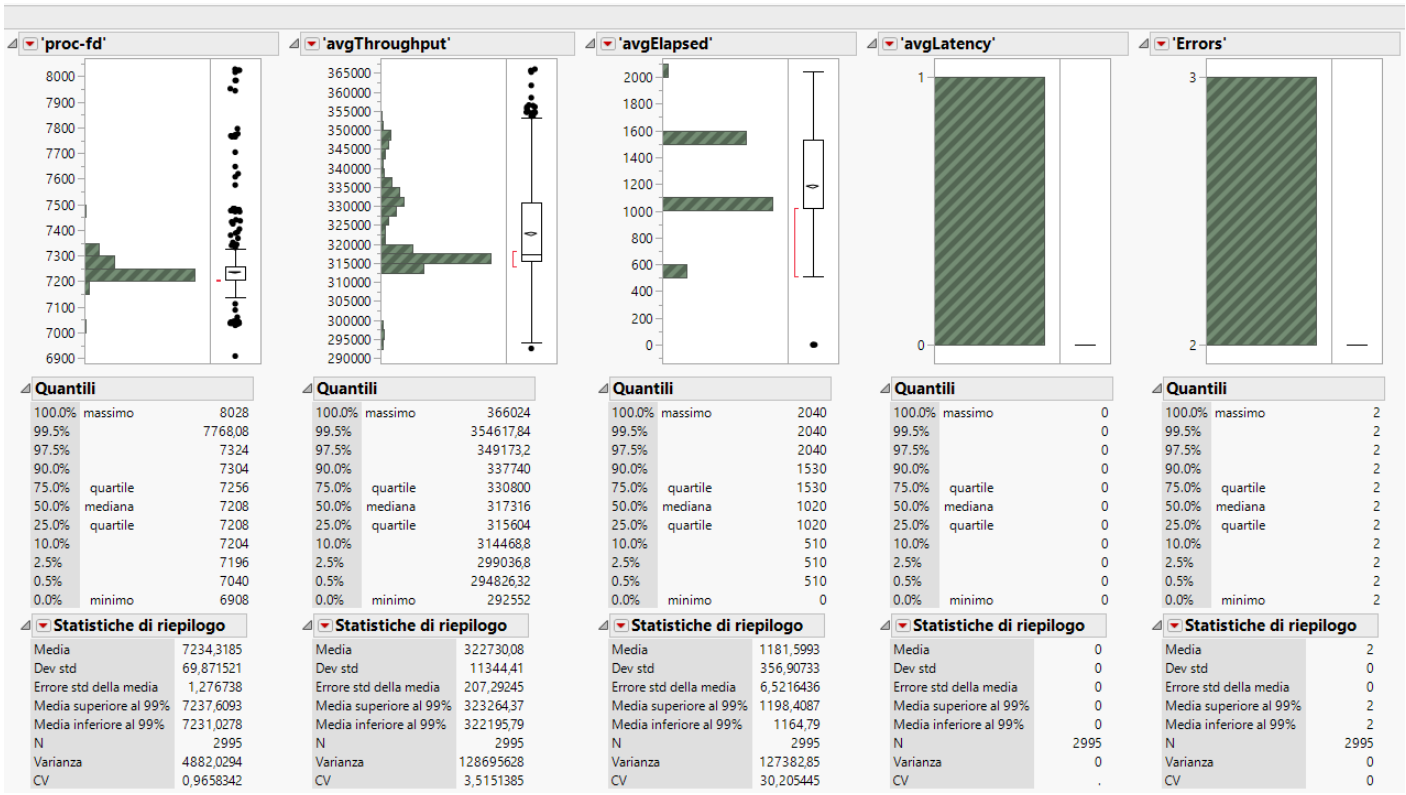

Figura 5: Distribuzioni delle feature

Come si può osservare dalle distribuzioni mostrate, gli attributi che presentano CV nullo sono: *active*, *anonpages*, *avglatency* ed *errors*.

Un'ulteriore scrematura delle feature si può effettuare osservando colonne che presentano stessi valori. Ad occhio, le colonne *memfree* e *writeback* risultano essere uguali. Per esserne certi, si è utilizzata la funzionalità del tool che permette di applicare delle formule sui dati delle colonne.

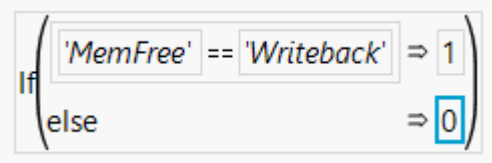

Figura 6: Formula confronto colonne

Del risultato prodotto, si è analizzata la distribuzione verificando che questa abbia coefficiente di variazione nullo.

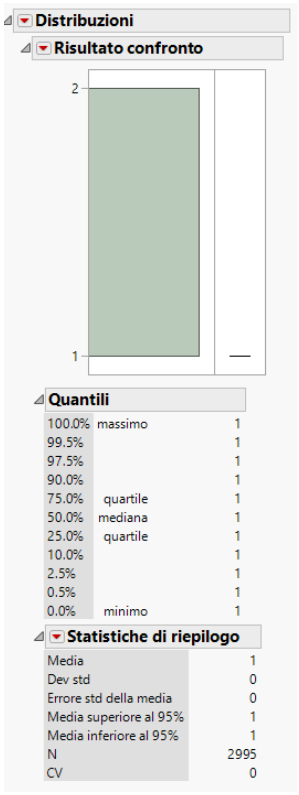

Figura 7: Distribuzioni del risultato del confronto

Il risultato di questa fase è una selezione di 19 feature a partire dalle 24 iniziali.

0.1.2 Principal Component Analysis

A questo punto si è proceduto ad effettuare un'analisi delle componenti principali, allo scopo di ridurre ulteriormente il numero di feature conservando la maggior parte della varianza. Ci si pone come obiettivo di conservare almeno il 95% della varianza totale.

Il risultato può essere osservato in forma grafica tramite uno score plot e loading plot.

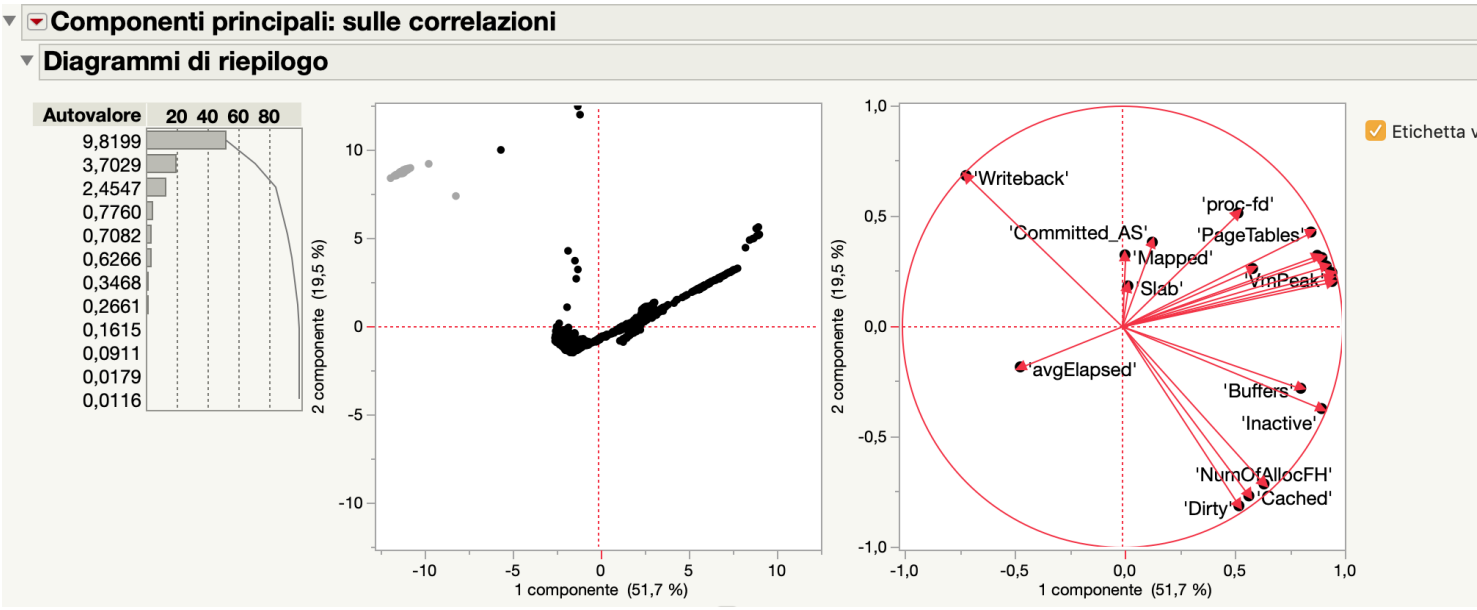

Figura 8: Score plot e loading plot generate dal tool

Il tool genera una vista degli autovalori della matrice di correlazione con la relativa percentuale di varianza per ogni componente ottenuta.

Autovalori

Numero	Autovalore	Percentuale	20 40 60 80	Percentuale cumulativa
1	10,3609	51,805		51,805
2	4,1387	20,694		72,498
3	2,4767	12,383		84,882
4	0,7760	3,880		88,762
5	0,7083	3,542		92,303
6	0,6267	3,134		95,437
7	0,3471	1,736		97,172
8	0,2661	1,331		98,503
9	0,1615	0,808		99,311
10	0,0913	0,456		99,767
11	0,0179	0,090		99,857
12	0,0116	0,058		99,915
13	0,0080	0,040		99,955
14	0,0061	0,031		99,985
15	0,0014	0,007		99,992
16	0,0009	0,005		99,997
17	0,0006	0,003		100,000
18	0,0000	0,000		100,000
19	0,0000	0,000		100,000

Figura 9: Autovalori della matrice di correlazione

Per rispondere all'obiettivo, si sceglie un numero di componenti principali pari a 6, rappresentativi del 95.437% della varianza totale.

0.1.3 Clustering

A valle della PCA effettuata, si vuole ridurre ulteriormente il dataset, con la differenza di voler diminuire il numero di istanze. A tale scopo si utilizza la tecnica del clustering di tipo gerarchico sulle componenti principali individuate, tramite la quale, scegliendo come metrica la distanza di Ward, si vuole individuare un trade-off tra la necessità di conservare una buona percentuale di varianza e quella di avere un numero accettabile di cluster.

Il risultato della fase di clustering è apprezzabile tramite il dendrogramma, prodotto dal tool.

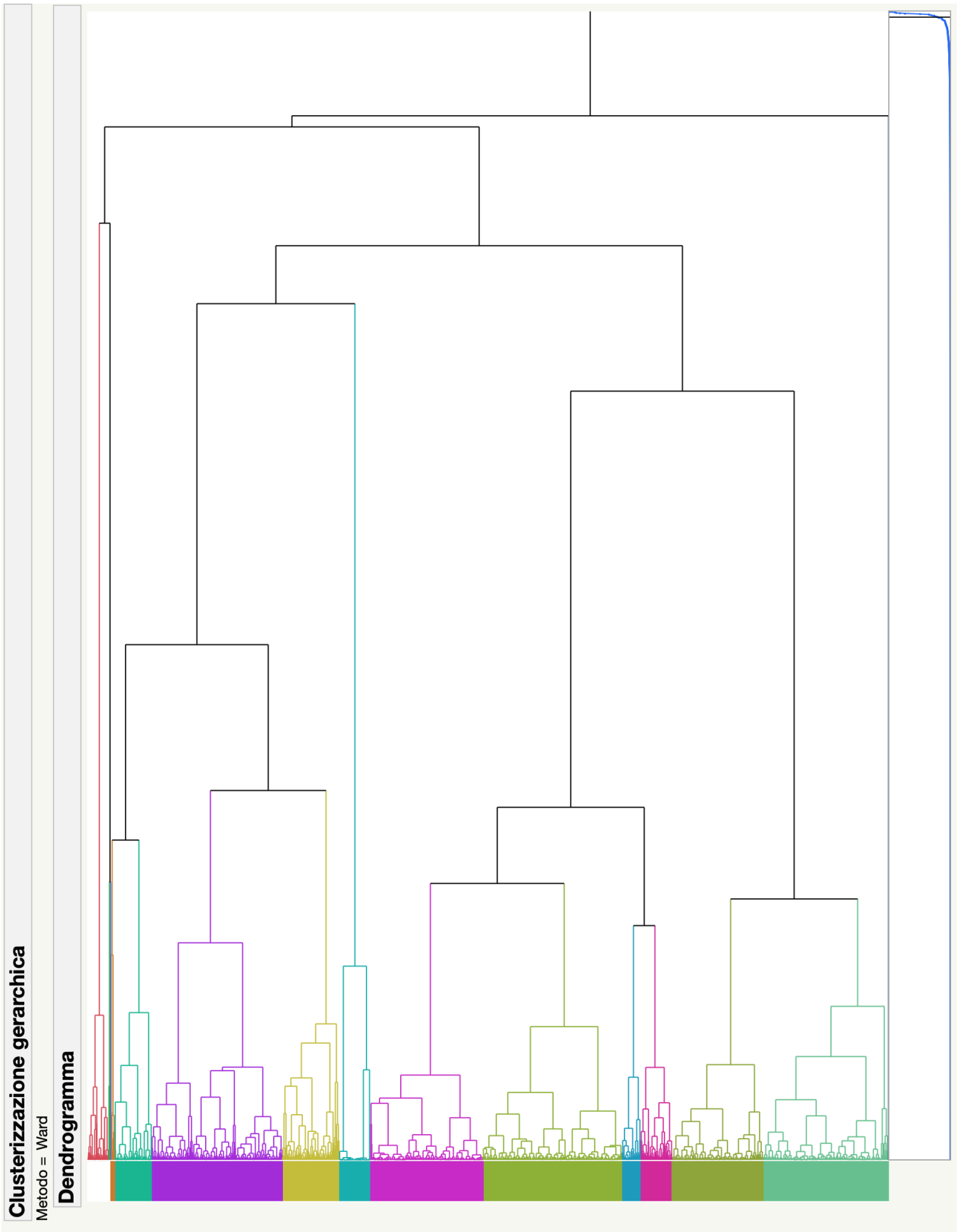

Figura 10: Dendrogramma

Inoltre il tool produce anche una tabella contenente, per ogni partizione, il relativo valore della distanza di Ward.

Cronologia di clusterizzazione

Numero di cluster	Distanza	Leader	Subordinato
30	4,65279127	113	379
29	4,70158610	996	1048
28	4,82475504	2057	2064
27	4,83888540	113	139
26	4,93013294	794	859
25	4,96878746	2185	2611
24	5,39058432	2186	2188
23	6,09832547	91	225
22	6,95244103	1064	1386
21	7,09509431	91	378
20	7,54600054	1	2
19	7,68446268	794	852
18	8,01290011	2186	2189
17	10,10841658	942	996
16	10,69609807	86	901
15	11,32656202	90	113
14	12,22359787	1114	2057
13	13,61384675	2185	2186
12	14,42223242	1058	1064
11	14,50888944	79	85
10	16,68472836	86	794
9	18,39417541	1058	1114
8	19,27064120	90	91
7	26,87885443	86	90
6	40,10315213	1058	2185
5	44,66565296	86	942
4	47,68943938	86	1058
3	48,86473683	1	79
2	53,88674745	1	86
1	54,46632673	1	2995

Figura 11: Cronologia di clusterizzazione

Al fine di scegliere il numero di cluster si valuta, per ogni partizione, la percentuale di varianza spiegata rispetto a quella ottenuta a valle della PCA. La varianza spiegata, in termini percentuali, è ottenuta tramite la formula:

$$V_{TOT} = V_{PCA} - (V_{PCA} * \frac{D_i}{D_t})$$

dove

- V_{PCA}è la percentuale di varianza spiegata a valle della PCA;
- D_iè la distanza di Ward della i-esima partizione di cluster;

- D_tè la distanza di Ward della partizione con un solo cluster.

Di seguito sono state calcolate alcune percentuali di varianza corrispondenti a partizioni significative, in quanto corrispondono a "salti" significativi di valori di distanza.

Numero di cluster	**Varianza spiegata (%)**
2	1.014
6	25.167
7	48.339
10	66.201
15	75.589
20	82.215
25	86.730

0.2 Conclusioni

Per avere un buon trade-off tra numero di cluster e varianza spiegata, si è scelto un numero di cluster pari a 15 corrispondente ad una varianza spiegata pari al 75.589%. Tramite una tecnica di campionamento casuale, è possibile costruire un workload sintetico, che si riporta di seguito.

'VmPeak'	'VmSize'	'VmHWM'	'VmRSS'	'VmPTE'	'Threads'	'MemFree'	'Buffers'	'Cached'	'Active'	'Inactive'	'Dirty'	'Writeback'	'AnonPages'	'Mapped'	'Slab'
152272	150216	25172	25104	160	47	5606032	31972	334648	0	141324	302672	5606032	8159224	132	0
153484	151436	26612	26472	160	14	5008324	83852	767660	0	380596	597076	5008324	8159224	108968	0
170344	170340	31144	31140	200	43	4471580	94884	1259568	0	597900	891332	4471580	8159224	459632	0
170344	170340	31916	31912	200	9	4699488	98052	1084804	0	386696	882460	4699488	8159224	107284	0
215352	210232	61784	57384	308	161	4531556	204212	1118440	0	523920	908436	4531556	8159224	320	0
197292	190944	43804	39220	276	99	4550072	203028	1118316	0	502284	910616	4550072	8159224	20	0
197292	190632	43804	38648	272	123	4551760	203040	1118316	0	501732	910612	4551760	8159224	20	0
215352	206136	61784	53204	312	9	4538840	204700	1118500	0	521332	907388	4538840	8159224	152	0
172388	170340	32192	31912	200	9	4663128	104912	1117060	0	408968	897220	4663128	8159224	32	0
172392	170340	33228	31628	200	15	4580048	185536	1117816	0	458488	928836	4580048	8159224	40	0
185964	177968	41280	35376	224	125	4564996	193932	1117980	0	475636	923976	4564996	8159224	144	0
186864	183260	41320	36660	248	79	4560980	195608	1118032	0	480616	922020	4560980	8159224	32	0
172392	170340	32696	31520	200	47	4621756	145184	1117288	0	420552	925784	4621756	8159224	164	0
172392	170340	32696	31340	200	9	4605088	161708	1117400	0	428596	934200	4605088	8159224	44	0
170344	170340	31144	31140	200	52	4655664	89480	1090572	0	541220	773332	4655664	8159224	324400	1760

Figura 12: Esempio di workload sintetico

'PageTables'	'Committed_AS'	'NumOfAllocFH'	'proc-fd'	'avgThroughput'	'avgElapsed'	'avgLatency'	'Errors'	Cluster
77368	23772	27884	7196	294908	1530	0	2	1
126120	28324	92116	7984	344944	2040	0	2	2
134504	29956	116216	8024	365692	1530	0	2	3
86152	24092	109392	7620	319664	1020	0	2	4
109728	23848	112956	7316	351436	510	0	2	5
91760	23848	114892	7284	338824	1020	0	2	6
90960	23848	113344	7276	337028	510	0	2	7
105548	23848	110196	7320	346020	1530	0	2	8
84216	23740	109024	7204	317680	1530	0	2	9
83972	23844	110996	7208	314256	1530	0	2	10
87656	23848	113756	7232	324712	2040	0	2	11
89076	23848	114416	7252	327404	1530	0	2	12
83864	23740	109872	7208	314000	1020	0	2	13
83684	23740	110308	7208	315100	1020	0	2	14
134500	29956	106788	8028	365264	2040	0	2	15

Figura 13: Esempio di workload sintetico