



UNIVERSITÀ DI PISA

PROGETTO PER IL CORSO DI DATA MINING
A.A. 2017/2018

Analisi del Dataset Human Resources Analytics

Gianmarco Ricciarelli
Maria Cristina Uccheddu
Stefano Carpita

Indice

1 Data Understanding	1
1.1 Obiettivi	1
1.2 Data semantics	1
1.3 Distribution of the variables and statistics	1
1.4 Data quality	4
1.5 Variable transformations	4
1.6 Pairwise correlations and eventual elimination of redundant variable	4
2 Clustering analysis	6
2.1 Clustering Analysis by K-means	6
2.1.1 Choice of attributes and distance function	6
2.1.2 Identification of the best value of k	6
2.1.3 Characterization of the obtained clusters	6
2.2 Hierarchical clustering	8
3 Association Rules	10
4 Classification	11

1 | Data Understanding

1.1 Obiettivi

In questo progetto viene analizzato il dataset (simulato) *Human Resources Analytics* contenente le informazioni sui dipendenti di un'azienda fittizia. Come mostrato in Figura 1.1 su un totale di 14999 dipendenti il 24%, corrispondente a 3571 lavoratori, ha lasciato l'azienda. Gli obiettivi primari dell'analisi sono i seguenti:

- capire i motivi principali per cui i lavoratori hanno lasciato l'azienda;
- predire probabilisticamente se un lavoratore lascerà in futuro l'azienda;
- indicare al management dell'azienda dei provvedimenti da attuare per ridurre il numero di impiegati che la abbandonano.

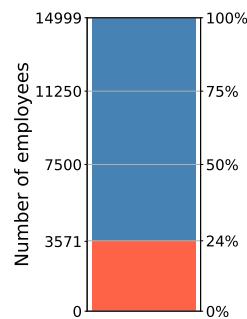


Figura 1.1: Numero di lavoratori

1.2 Data semantics

Il dataset è composto da 10 variabili relative ai dipendenti dell'azienda, riportate in tabella 1.1, delle quali 5 sono di tipologia categorica, di cui una ordinale, e 5 di tipologia numerica.

La variabile *Left* suddivide il dataset tra i dipendenti che hanno lasciato l'azienda e quelli che ci lavorano attualmente, associando alle rispettive categorie i valori 1 e 0. I dipendenti lavorano in 10 diversi dipartimenti indicati nella variabile *Department*, che è stata rinominata rispetto all'originale *Sales* per chiarezza semantica. La promozione o meno di un dipendente durante gli ultimi 5 anni è espressa dalla variabile *Promotion last 5 Year* con un 1 in caso positivo e con 0 altrimenti. *Work Accident* indica con un 1 il coinvolgimento di un dipendente in un incidente sul lavoro, e con 0 il caso contrario. *Salary* esprime il livello (*low, medium, high*) nel quale rientra il salario del dipendente. Con la variabile *Satisfaction Level* viene fornita una valutazione quantitativa del livello di soddisfazione di ciascun dipendente, in un range che va da un valore minimo di 0 ad un massimo di 1. *Last Evaluation* fornisce l'ultima valutazione riguardo le performance del dipendente, compresa tra 0 ed 1. *Average Montly Hours* rappresenta la media delle ore di lavoro in un mese mentre *Time Spend Company* corrisponde al numero di anni trascorsi dal dipendente all'interno dell'azienda. *Number Projects* riporta il numero di progetti completati da ciascun dipendente durante il periodo di lavoro. Le descrizioni delle variabili sono state estrapolate dai metadati forniti assieme al dataset sulla pagina di Kaggle¹ nella quale il dataset è pubblicato.

Variable	Type
Left	categorical
Department	categorical
Promotion_last_5years	categorical
Work_accident	categorical
Salary	ordinal
Satisfaction_level	continuous
Last_evaluation	continuous
Average_montly_hours	discrete
Time_spend_company	discrete
Number_projects	discrete

Tabella 1.1: Variabili presenti nel Dataset e rispettivi tipi.

1.3 Distribution of the variables and statistics

In questo paragrafo vengono presentati i grafici relativi alla distribuzione dei valori assunti dalle variabili descritte nella sezione precedente. Per dare una migliore interpretazione a questi abbiamo deciso di sostenere un'analisi accurata che contraddistingue i dipendenti che lavorano nell'azienda, rappresentati dal colore blu, e quelli che invece la hanno lasciata, rappresentati dal colore rosso.

Prima di tutto vogliamo studiare la distribuzione dei dipendenti rispetto alle variabili categoriche escludendo la distribuzione dei dipendenti rispetto a left in quanto già esplicata in precedenza nella sezione Obiettivi.

¹<https://www.kaggle.com/>

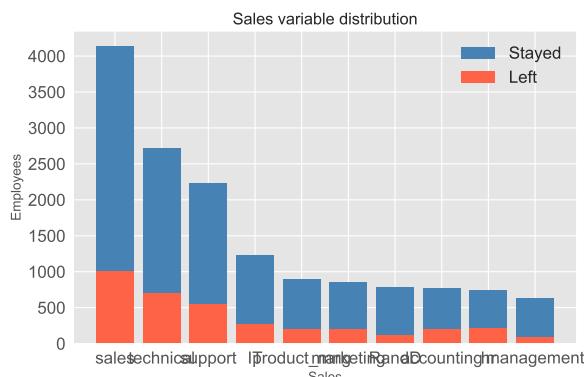


Figura 1.2: Distribuzione relativa alla variabile Sales

Dipartimento	InAzienda	OutAzienda	TotDip
1 Sales	3126	1014	4140
2 technical	2023	697	2720
3 support	1674	555	2229
4 IT	954	273	1227
5 productMng	704	198	902
6 marketing	655	203	858
7 RandD	666	121	787
8 accounting	563	204	767
9 hr	524	215	739
10 management	539	91	630
Totale	11428	3571	14999

Tabella 1.2: Distribuzione salario per dipendente.



Figura 1.3: Distribuzione relativa alla variabile Work Accident

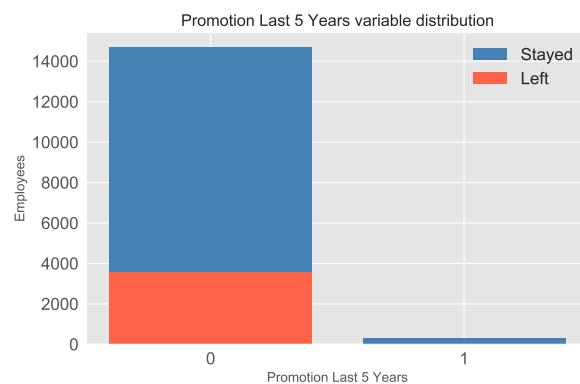


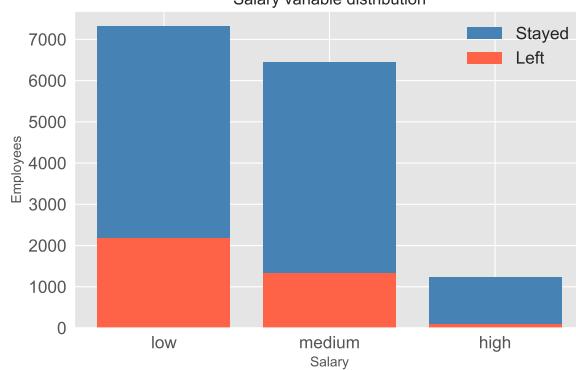
Figura 1.4: Distribuzione relativa alla variabile Promotion Last 5 Years

Nel grafico a sinistra si studia il rapporto tra i dipendenti e la presenza o meno di un infortunio durante il periodo di lavoro all'interno dell'azienda e si è riscontrato che di quelli che l'hanno lasciata soltanto 169 impiegati hanno avuto un incidente sul lavoro (circa il 4,75% su 3571 e circa il 1,13% dei dipendenti totali), mentre gli impiegati ancora all'interno dell'azienda ad aver subito un incidente sono 2000 (circa il 17,5% su 11428 e circa il 13,35% dei dipendenti totali)

Nel grafico a destra invece riportiamo ciascun dipendente al fatto che questo sia stato promosso negli ultimi 5 anni oppure no, possiamo ricavare una informazione importante, la gran parte degli impiegati che hanno lasciato l'azienda non ha avuto una promozione negli ultimi 5 anni a parte 19 impiegati che è stata promossa (circa il 0.5%, invece circa il 0,13% dei dipendenti totali) , praticamente impercettibili alla vista del grafico. Degli impiegati rimasti in 300 hanno ottenuto una promozione su 11428 di quelli rimasti sono stati promossi (circa il 2,62% di quelli rimasti e circa il 2% dei dipendenti totali).

Una volta studiate le distribuzioni categoriche continuiamo l'analisi con gli altri attributi.

Partiamo dalla distribuzione del salario:



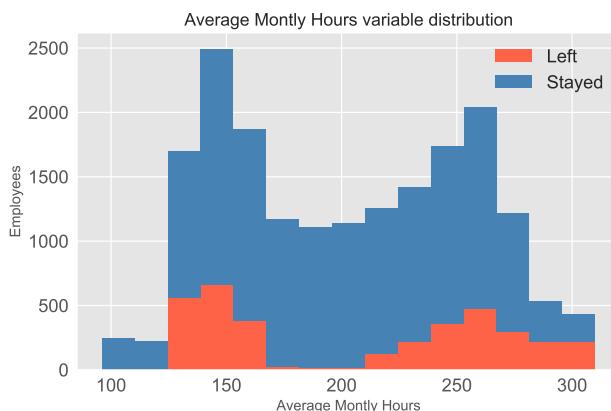
Dipendente	InAzienda	OutAzienda
Low	5144 (~45.02%)	2172 (~60.8%)
Medium	5129 (~44.88%)	1317 (~36.9%)
High	1155 (~10.1%)	82 (~2.3%)

Tabella 1.3: Distribuzione salario per dipendente.

Figura 1.5: Distribuzione relativa alla variabile Salary

Le percentuali che vengono indicate nella tabella non sono in base alla totalità dei dipendenti ma riguardano solo il tipo di dipendente definito dalla colonna di appartenenza.

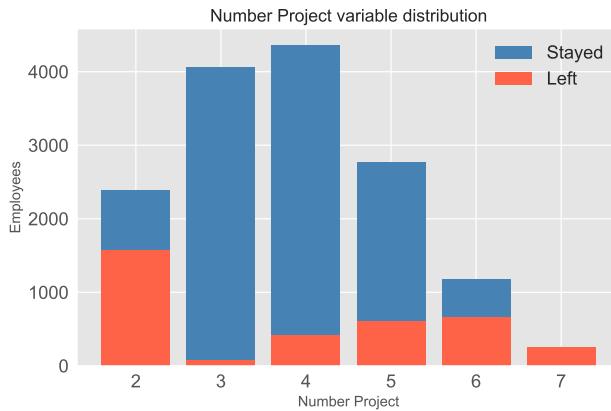
Proseguendo poi con il rapporto tra dipendenti ancora in azienda e non, e il numero delle ore di lavoro in media:



Dipendente	InAzienda	OutAzienda
Media	199.06	207.42
Dev.std.	45.68	61.20
min,max	96 , 287	126 , 310

Figura 1.6: Distribuzione relativa alla variabile *Average Monthly Hours*

Dipendenti left e non rispetto al Numero di progetti:



Dipartimento	InAzienda	OutAzienda
2	821	1567
3	3983	72
4	3956	409
5	2149	612
6	519	655
7	0	256

Figura 1.7: Distribuzione relativa alla variabile *Number Project*

A questo punto è giusto analizzare il livello di soddisfazione dei dipendenti presente e quello della ultima valutazione:

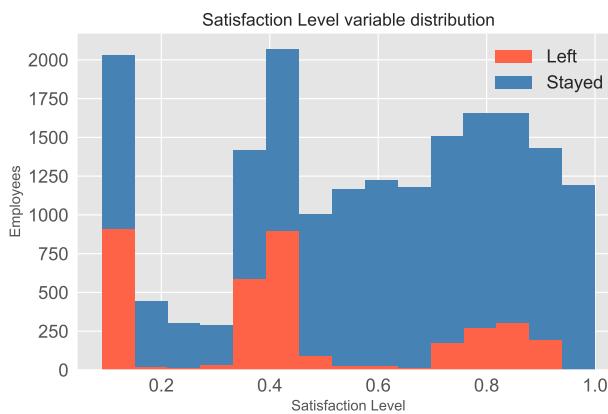


Figura 1.8: Distribuzione relativa alla variabile *Satisfaction Level*

Dipendente	InAzienda	OutAzienda
Media	0.67	0.44
Dev.std.	0.22	0.26
min,max	0.12 , 1.0	0.09 , 0.92

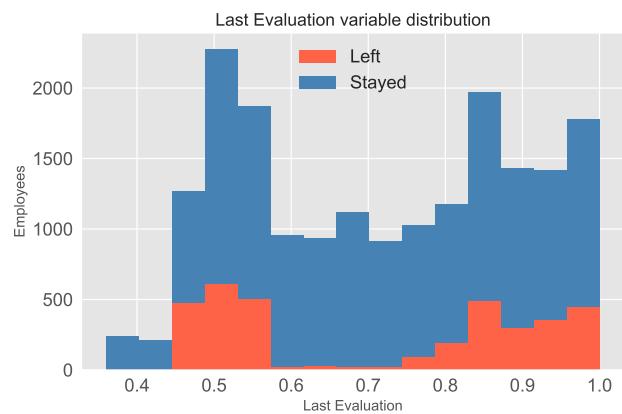


Figura 1.9: Distribuzione relativa alla variabile *Last Evaluation*

Dipendente	InAzienda	OutAzienda
Media	0.71	0.71
Dev.std.	0.16	0.19
min,max	0.36 , 1.0	0.45 , 1.0

1.4 Data quality

Nel dataset non risultano presenti nè missing values nè errori sintattici. L'individuazione dei possibili outliers delle variabili numeriche consiste nel verificare se siano presenti dei valori estremi rispetto alle relative distribuzioni. Un metodo robusto e di immediata applicazione è quello di rappresentare i dati tramite boxplot, identificando come outliers i valori che si trovano al di fuori dei whiskers, di lunghezza massima 1.5 IQR . Per una distribuzione gaussiana il numero di outliers attesi al di fuori dei whiskers è circa il 7%. In Fig. 1.10 si osserva che si hanno outliers solo per la variabile *Time Spend Company*, la cui distribuzione risulta però essere asimmetrica e con una coda destra ampia come mostrato anche in Fig. REF. Gli outliers individuati non sono quindi considerati anomali, ma caratteristici della distribuzione, e non vengono eliminati dal dataset nelle successive analisi. (Altri test comunemente utilizzati, come il test di Grubb o il criterio di Chauvenet, sono basati sull'assunzione di una distribuzione di probabilità gaussiana, che non si osserva per le variabili numeriche del dataset analizzato (spiegare)).

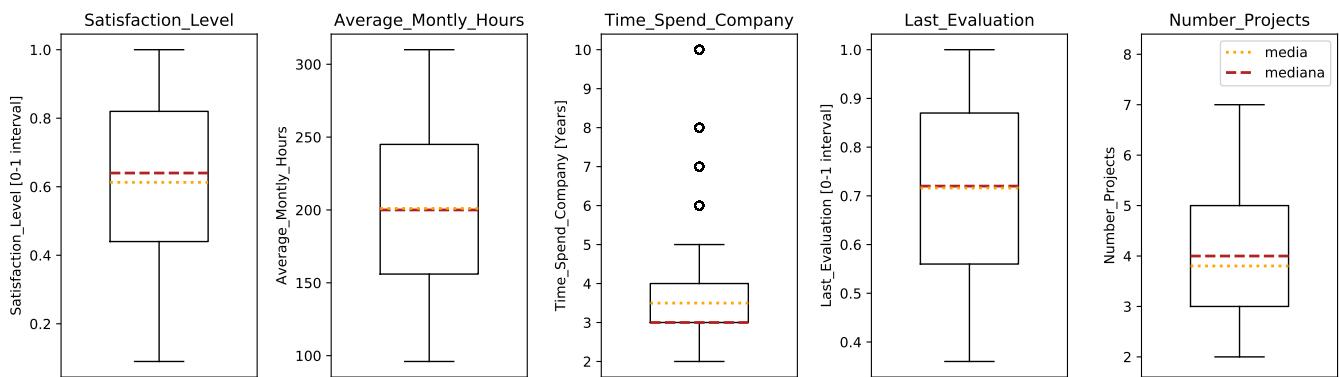


Figura 1.10: Boxplots per le variabili numeriche

Tramite k-nearest-neighbours mi pare che ci siano parecchi dati ripetutie, approfondire

1.5 Variable transformations

Analizzando il significato delle variabili presenti nel dataset, abbiamo deciso di rappresentare *Work Accident* e *Left* utilizzando il tipo categorico piuttosto che quello discreto. Questa scelta è stata motivata dall'analisi semantica delle due variabili, le quali forniscono una risposta del tipo "Sì o No" alle domande relative agli incidenti sul lavoro e all'abbandono o meno dell'azienda da parte dei dipendenti.

1.6 Pairwise correlations and eventual elimination of redundant variable

In questa sezione abbiamo studiato la correlazione ovvero la relazione lineare tra i vari attributi continuous o discreti. Dalla matrice riportata in Figura 1.11 possiamo rilevare se ci sia una correlazione positiva, nulla o negativa. Sia per quanto riguarda la correlazione positiva sia per quella negativa si caratterizzano in settori: con valori da 0 a 0.3 correlato debolmente, da 0.3 a 0.7 moderatamente o maggiore di 0.7 fortemente (rispettivamente per la negativa i segni saranno negativi). Da questo possiamo definire che ad avere una correlazione debole è la variabile *time_spend_company* con *left*, *last_evaluation*, *number_project* e *average_monthly_hours*. Queste ultime, ad eccezione di *left*, invece sono correlate fra loro in modo moderato con un valore massimo di 0.42 tra *average_monthly_hours* e *number_project*. Il valore 1 indica la correlazione con se stesso che infatti è massima. Dal punto di vista della correlazione negativa, abbiamo debolmente correlati *left* con *work_accident* e *satisfaction_level* con *number_projects* e *time_spend_company*. Abbiamo invece una correlazione negativa moderata tra *left* e *satisfaction_level* di valore -0.39.

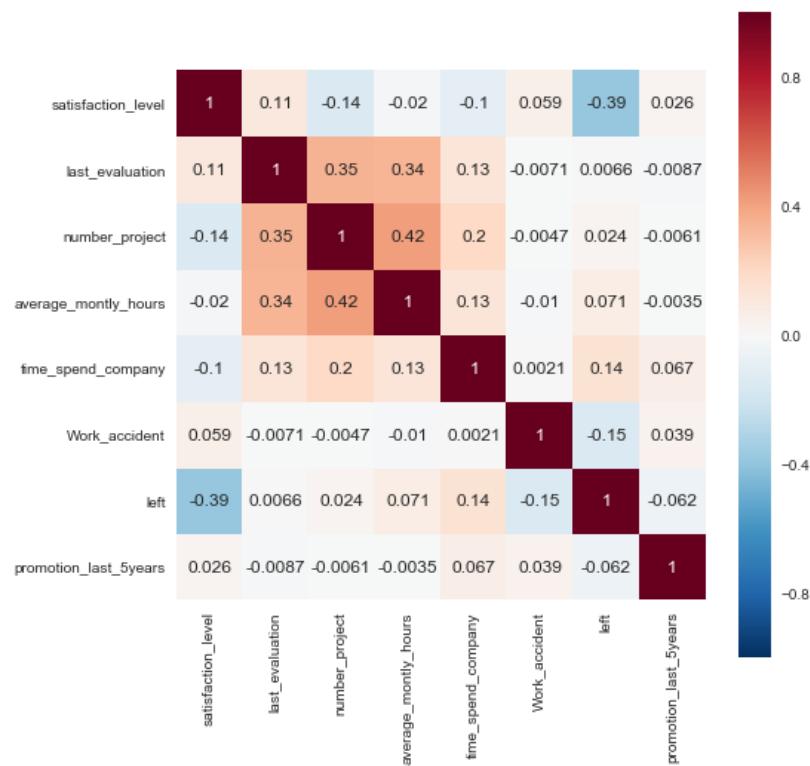


Figura 1.11: Correlation Matrix delle variabili presenti nel Dataset.

2 | Clustering analysis

2.1 Clustering Analysis by K-means

2.1.1 Choice of attributes and distance function

Le variabili sulle quali abbiamo deciso di applicare la Cluster Analysis tramite K-means sono le due variabili di tipo continuous presenti nel Dataset, ossia *Satisfaction Level* e *Last Evaluation*. Le variabili di tipo categorico sono state scartate al momento della scelta dato che la natura stessa dell'algoritmo prevede il suo utilizzo su variabili di tipo numerico.

Approfondire su variabili discrete

Nell'implementazione dell'algoritmo da noi utilizzata è stato deciso di applicare la distanza euclidea come distance function.

2.1.2 Identification of the best value of k

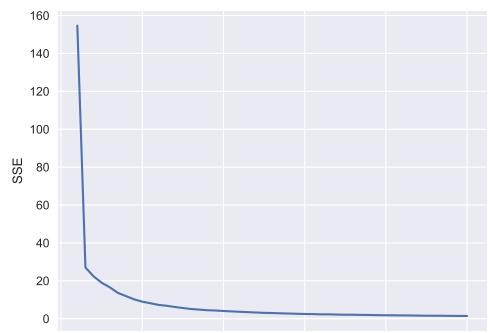
Al fine di identificare il miglior numero k di clusters da utilizzare, abbiamo tenuto conto dell>Error Sum of Squares (SSE) per ogni iterazione dell'algoritmo, svolta a partire da un valore iniziale di k pari a 2 fino ad un valore massimo di 50. Rappresentato in Figura 2.1a troviamo l'andamento dell'SSE per i cluster relativi ai dipendenti che hanno lasciato l'azienda. Possiamo notare il valore ottimale di k pari a 3, che è la posizione sull'asse dei cluster dove la curva inizia il suo percorso discendente. In figura 2.1b possiamo vedere la stessa cosa, ma per i dipendenti che sono rimasti all'interno dell'azienda. In questo caso notiamo il valore ottimale di k pari a 5.

2.1.3 Characterization of the obtained clusters

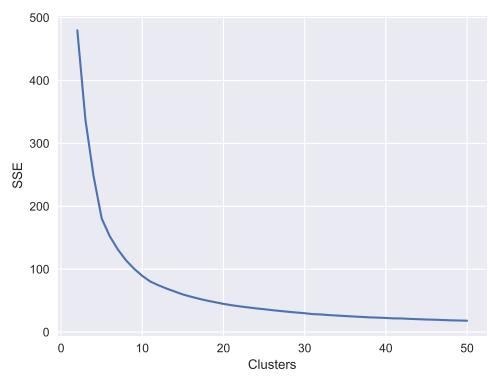
Andiamo adesso a dare una descrizione dei cluster emersi dall'applicazione dell'algoritmo K-means con i parametri che abbiamo deciso di utilizzare. Cominciando con l'osservare lo scatterplot relativo ai cluster degli impiegati che hanno lasciato l'azienda, Figura , notiamo che sono ben visibili 3 categorie distinte di impiegati, che possiamo descrivere discorsivamente come

- Impiegati con un livello di soddisfazione basso e un alto score nella valutazione. Questo cluster ci suggerisce la motivazione per cui questi impiegati hanno lasciato l'azienda, ossia una quantità troppo elevata di ore di lavoro, dalla quale deriva probabilmente l'alto score nella valutazione.
- Impiegati con un basso livello di soddisfazione, compreso tra 0.3 e 0.5, e un altrettanto basso score nella valutazione, compreso tra 0.2 e 0.55. La scarsa produttività e soddisfazione suggeriscono che questo gruppo di impiegati ha deciso di lasciare l'azienda per insoddisfazione verso la posizione lavorativa offerta.
- Impiegati con un alto livello di soddisfazione e un alto score nella valutazione, che probabilmente hanno lasciato l'azienda dopo aver ricevuto un'offerta di lavoro più vantaggiosa.

Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum.



(a)



(b)

Figura 2.1: Nella Figura 2.1a troviamo la rappresentazione della curva con cui l'SSE dei cluster relativi ai dipendenti che hanno lasciato l'azienda decresce con l'aumentare del numero di cluster . Lo stesso vale per la Figura 2.1b, ma per i dipendenti che sono ancora all'interno dell'azienda.

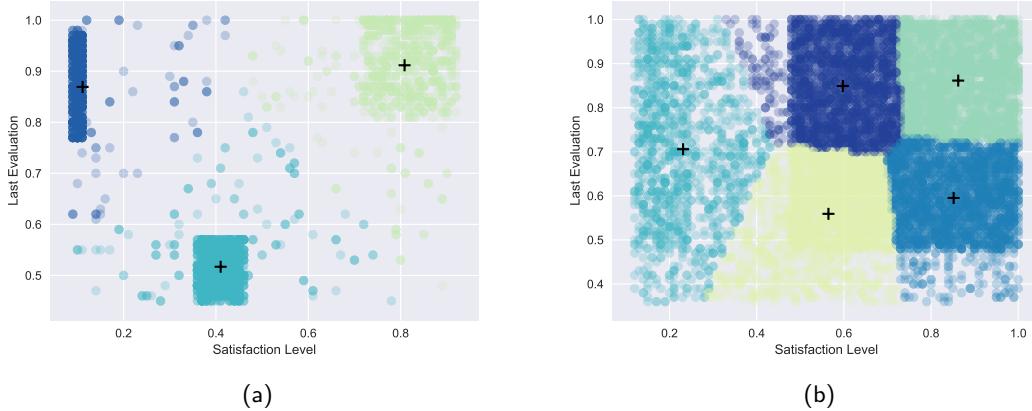


Figura 2.2: Nella Figura 2.2a troviamo i cluster emersi dall'analisi dei dipendenti che hanno lasciato l'azienda, nella Figura 2.2b troviamo invece quelli emersi dall'analisi dei dipendenti rimasti nell'azienda.

Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetur adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetur.

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

2.2 Hierarchical clustering

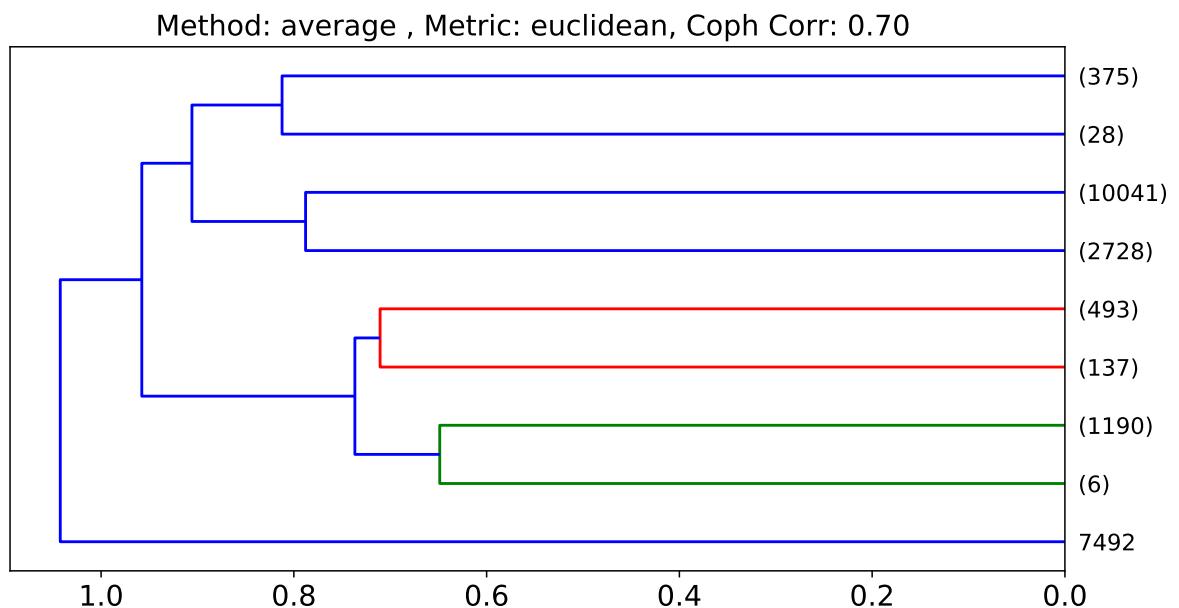


Figura 2.3: Dendrogramma per method X e metrica Y

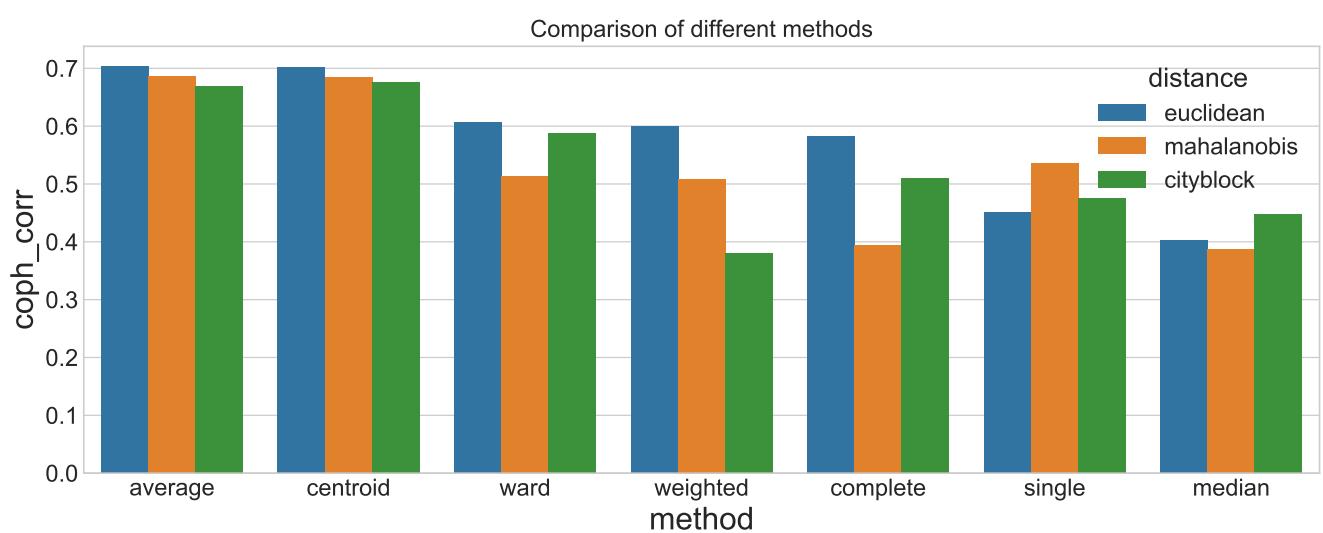


Figura 2.4: Confronto tra diversi metodi

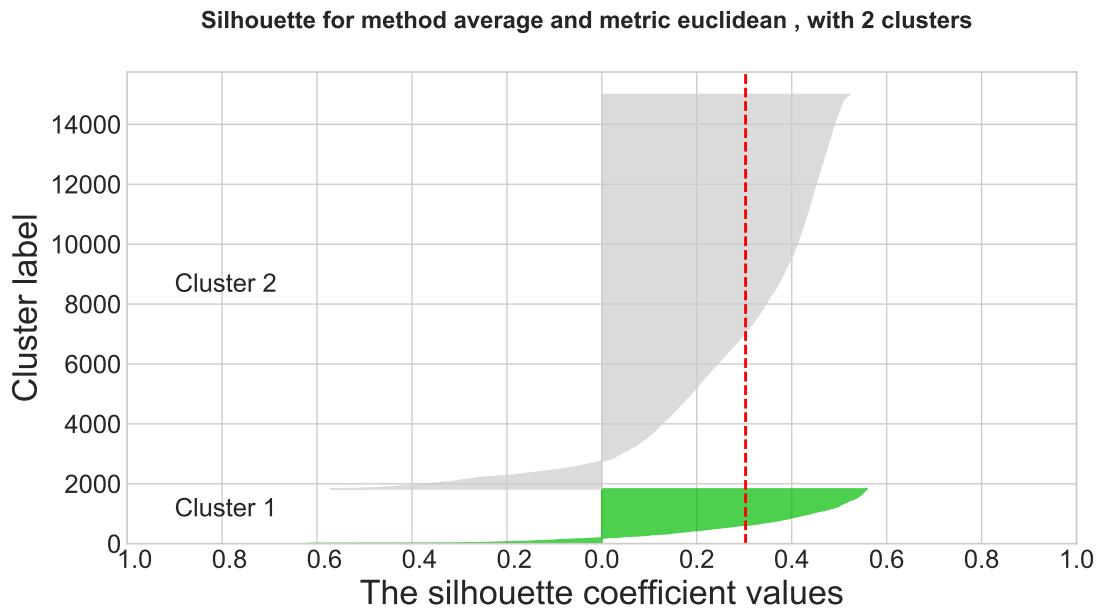


Figura 2.5: Confronto tra silhouette medie, per due clusters

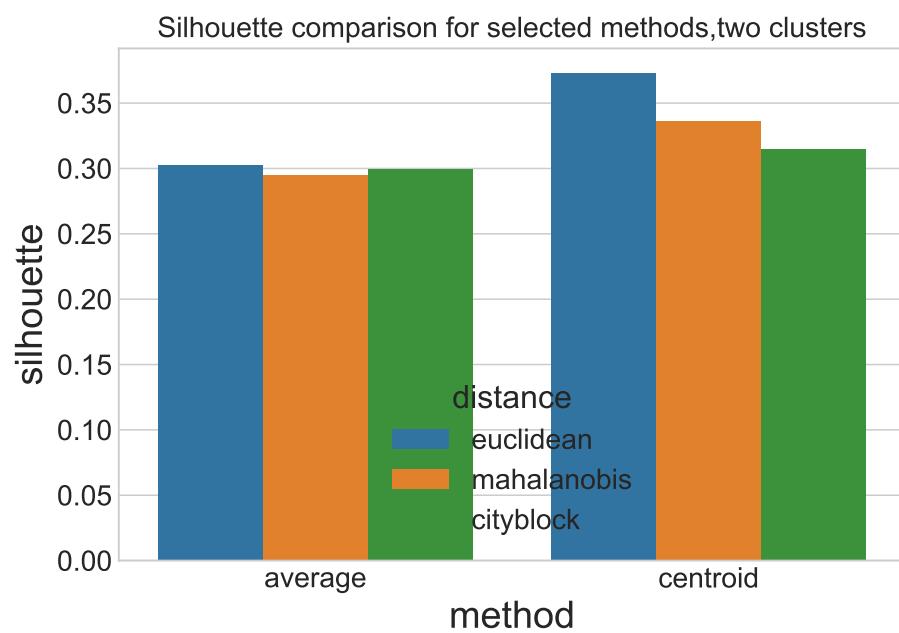


Figura 2.6: Confronto tra silhouette medie, per due clusters

3 | Association Rules

4 | Classification