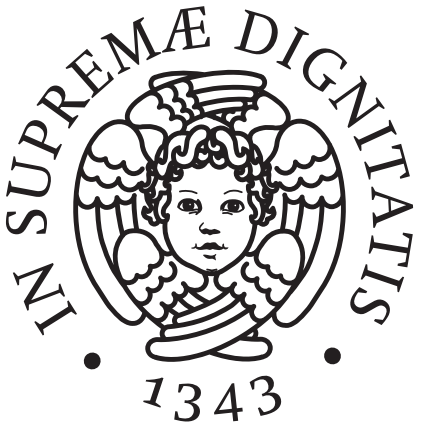


Università di Pisa

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 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.

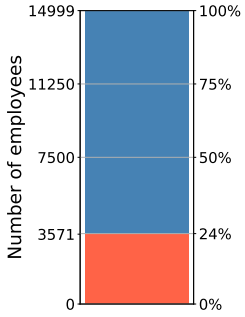


Figura 1.1: Numero di lavoratori

1.2 Data semantics

In tabella 1.1 sono riportate le variabili relative ai dipendenti dell'azienda e la corrispondente tipologia.

Con la prima 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 delle performance del dipendente. *Number Project* riporta il numero di progetti completati da ciascun dipendente durante il periodo di lavoro. *Average Montly Hours* rappresenta la media delle ore di lavoro in un mese. *Time Spend Company* corrisponde al numero di anni trascorsi dal dipendente all'interno dell'azienda. *Work Accident* esprime con un 1 il coinvolgimento di un dipendente in un incidente sul lavoro, altrimenti viene impostato come 0. *Left* è utilizzata per tener traccia dei dipendenti che hanno lasciato l'azienda, per i quali viene usato il valore 1, mentre per quelli che sono rimasti viene usato il valore 0. *Promotion last 5 Year* esprime con un 1 se il dipendente è stato promosso negli ultimi 5 anni, altrimenti assume il valore 0. *Sales* definisce il dipartimento nel quale il dipendente lavora, e *Salary* esprime il livello (*low*, *medium*, *high*) nel quale rientra il salario del dipendente. Le descrizioni delle variabili sono state estrapolate dai metadati forniti assieme al Dataset sulla pagina di Kaggle [1] nella quale il Dataset è contenuto. La scelta di catalogare le colonne *Work Accident*, *Left* e *Promotion Last 5 Years* come variabili di tipo categorico verrà opportunamente motivata nelle s ezioni successive.

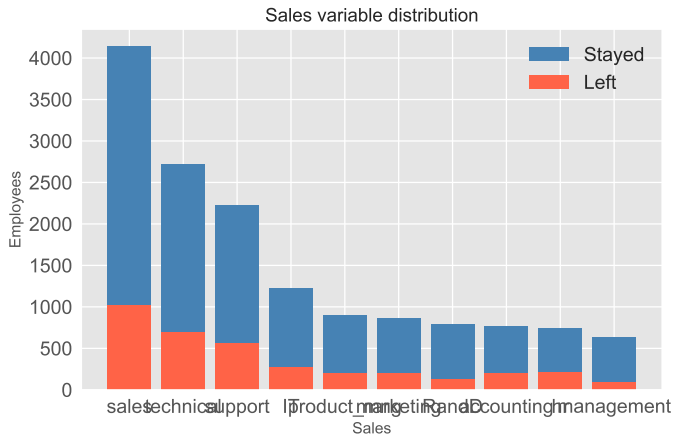
Field	Type
satisfaction_level	continuous
last_evaluation	continuous
number_project	discrete
average_montly_hours	discrete
time_spend_company	discrete
salary	ordinal
work_accident	categorical
promotion_last_5years	categorical
sales	categorical
left	categorical

Tabella 1.1: Variabili presenti nel Dataset e rispettivi tipi.

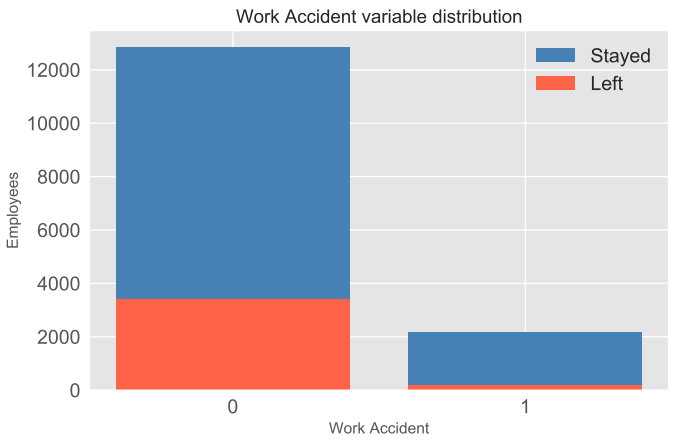
1.3 Distribution of the variables and statistics

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. Come si può constatare dal grafico in Figura 1.1 in un totale di 14999 dipendenti analizzati: 3571 di questi hanno lasciato l'azienda (quasi il 24% del totale), mentre i restanti 11428 dipendenti (quasi il 76%) sono rimasti nella azienda. Prima di tutto vogliamo studiare la distribuzione dei dipendenti rispetto alle variabili categoriche. La distribuzione mostrata in Figura 1.2a indica la distribuzione degli impiegati nei vari settori: (parte mancante).
Nel grafico in Figura 1.2b 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 soltanto 169 impiegati tra i 3571 che se ne sono andati hanno avuto un incidente sul lavoro, mentre gli impiegati ancora all'interno dell'azienda ad aver subito un incidente sono 2000 su 11428.
Nel grafico in Figura 1.2c invece rapportiamo ciascun dipendente al fatto che questo sia stato promosso negli ultimi 5 anni oppure no, possiamo ricavare una informazione importante, la totalità degli impiegati che hanno lasciato l'azienda non ha avuto una promozione negli ultimi 5 anni. Soltato 19 impiegati su 11428 di quelli rimasti sono stati promossi. Una volta studiate le distribuzioni categoriche continuiamo l'analisi con gli altri attributi, per renderle più esplicative abbiamo affiancato a ciascun grafico la tabella che lo descrive in modo accurato e in modo che fosse autoesplicativo.

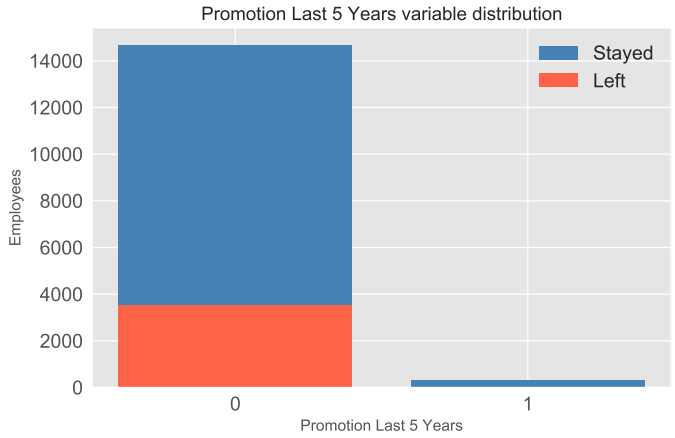
[1] https://www.kaggle.com/



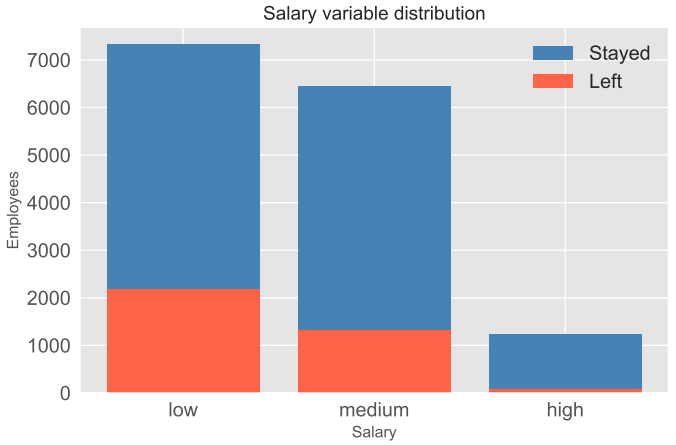
(a) Distribuzione relativa alla variabile *Sales*



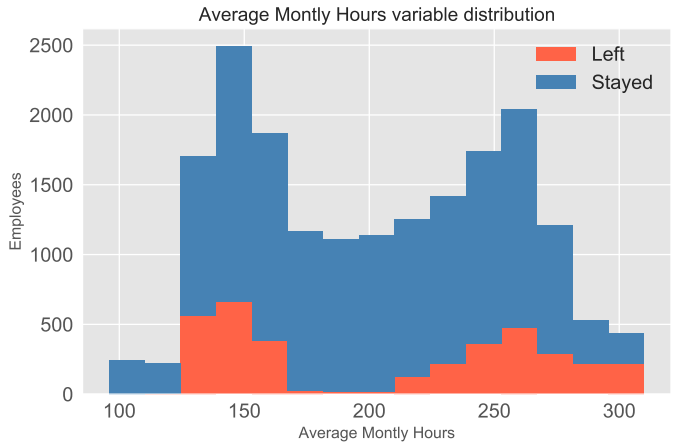
(b) Distribuzione relativa alla variabile *Work Accident*



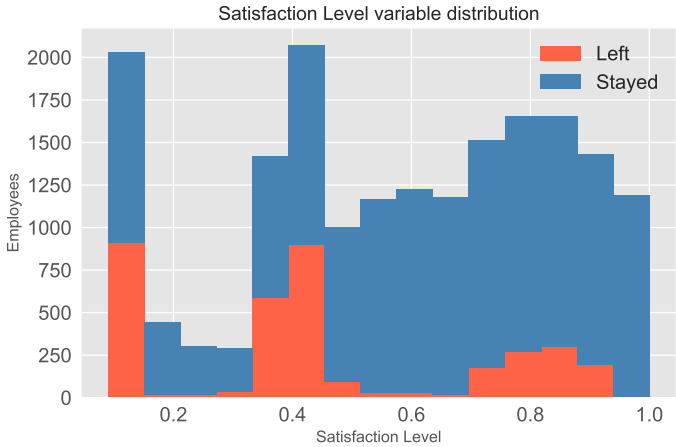
(c) Distribuzione relativa alla variabile *Promotion Last 5 Years*



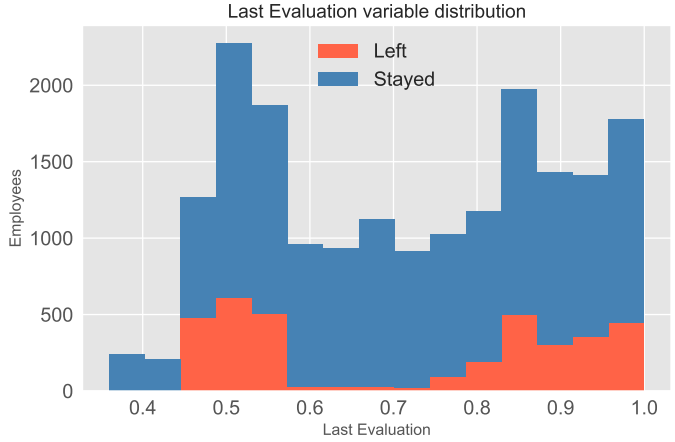
(d) Distribuzione relativa alla variabile *Salary*.



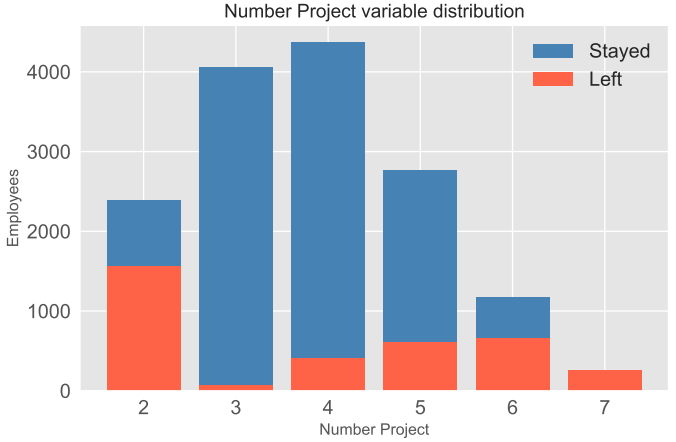
(e) Distribuzione relativa alla variabile *Average Montly Hours*



(f) Distribuzione relativa alla variabile *Satisfaction Level*



(g) Distribuzione relativa alla variabile *Last Evaluation*



(h) Distribuzione relativa alla variabile *Number Project*

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

(a) Distribuzione salario per dipendente.

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

(b) Il tempo medio di ore di lavoro al mese per dipendente.

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

(c) Livello di soddisfazione per il dipendente.

Dipendente	media	dev. std.	min,max
InAzienda	0.71	0.16	0.36 , 1.0
OutAzienda	0.71	0.19	0.45 , 1.0

(d) Livello di soddisfazione per il dipendente relativo all'ultima valutazione.

Abbiamo rapportato in primis i dipendenti con il proprio salario: possiamo osservare la distribuzione della variabile in Figura 1.2d e le relative percentuali in Tabella 1.2a. Le percentuali che vengono indicate nella tabella non sono in base alla totalità dei dipendenti ma riguardano solo il tipo di dipendente definito dalla riga di appartenenza.
Proseguendo poi con il rapporto tra dipendenti ancora in azienda e non, e il numero delle ore di lavoro in media, possiamo osservare la distribuzione della variabile in Figura 1.2e, e le relative percentuali riportate in Tabella 1.2b.
A questo punto è giusto analizzare il livello di soddisfazione dei dipendenti, del quale possiamo osservare la distribuzione in Figura 1.2f e le relative percentuali in Tabella 1.2c.
Possiamo rapportare rispetto alla precedente valutazione del livello di soddisfazione (cambiare valori della tabella) osservan

1.4 Data quality

- Missing values

- Outliers

L'individuazione dei possibili outliers di una variabile numerica consiste nel verificare se siano presenti dei valori estremi rispetto alla distribuzione dei dati osservati. I 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 in distribution of the variables). Un metodo robusto e di immediata applicazione è quello di osservare il boxplot dei dati, identificando come candidati outliers i valori che si trovano al di fuori dei whiskers, ovvero valori x della variabile osservata per cui $|x - \tilde{x}| > 2\,IQR(x)$, dove $\tilde{x}$ è la mediana ed $IQR(x)$ lo scarto interquartile.

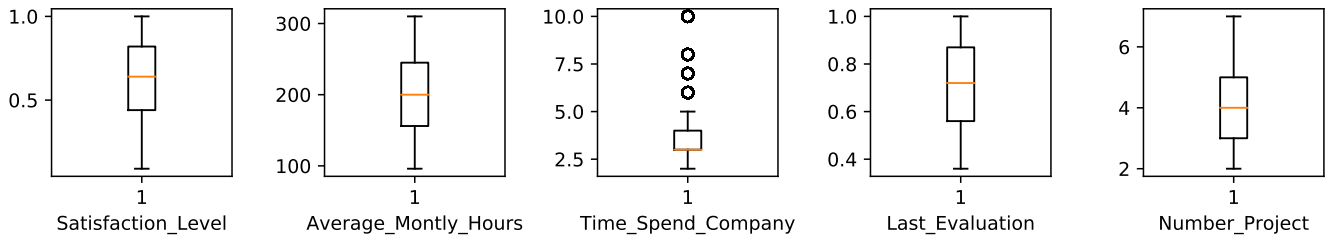


Figura 1.3: Boxplots per le variabili numeriche

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 1.4 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_projecy e time_spend_company. Abbiamo invece una correlazione negativa moderata tra left e satisfaction_level di valore -0.39.

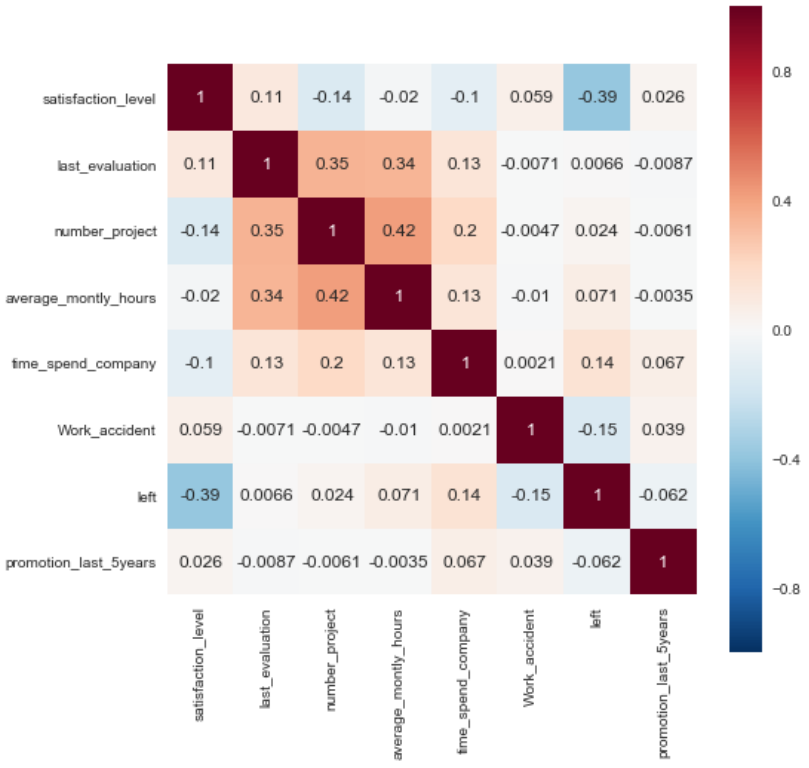


Figura 1.4: Correlation Matrix delle variabili presenti nel Dataset.

2 | Clustering analysis

In generale adotterei un approccio DRY, ovvero eviterei ripetizioni, dicendo ad esempio in questo paragrafo facciamo questo. (Modificare anche paragrafi precedenti, inoltre c'è il titolo che indica cosa si fa in ciascun punto). Eviterei di annunciare quello che verrà fatto per poi spiegarlo, spiegare direttamente nel punto giusto

La ricerca di gruppi di dipendenti con caratteristiche affini all'interno del dataset è stata eseguita utilizzando differenti tecniche di clustering. Per eseguire l'analisi sono state selezionate solamente le 5 variabili numeriche in Tab. 1.1, in modo da calcolare le distanze tra i dati in modo appropriato. Le variabili sono state standardizzate o normalizzate? Decidere spiegare confrontare. Come già specificato nella Sezione 1.5, i valori delle variabili discrete sono stati normalizzati in un intervallo compreso tra 0 e 1, al fine di rendere più agevole il confronto in fase di clustering.

2.1 Clustering Analysis by K-means

2.1.1 Choice of attributes and distance function

Scelta delle variabili vale per tutte e tre le tecniche, rimuovere da qua.

Le *variabili* sulle quali abbiamo deciso di applicare l'analisi tramite K-means sono un sottoinsieme di quelle a disposizione nel data set, in particolare, sono le variabili di tipo continuo *Satisfaction Level* e *Latest Evaluation*, e le variabili di tipo discreto *Number Project*, *Average Montly Hours* e *Time Spend Company*. Sono state escluse le variabili binarie. Vista la natura delle variabili utilizzate nell'applicazione dell'algoritmo, la distance function da noi utilizzata per quantificare la distanza tra due data objects è la *distanza Euclidea*, rappresentata dalla nota formula

Per quanto riguarda le formule o le mettiamo tutte o non ne mettiamo nessuna. Poichè non è richiesta la spiegazione degli algoritmi e poichè le quantità sono note alle persone che leggono eliminerei TUTTE le formule, concentriamoci sui risultati

$$dist(p,\ q)\ =\ \sqrt{\sum_{i=0}^{k}(q_i - p_i)^2}\quad ,$$

dove con q_i e p_i vengono rappresentati li i-esimi attributi dei data objects p e q. Infine, come ulteriori casi di studio, sono riportate anche le analisi relative alla sola porzione del data set contenente gli impiegati che hanno lasciato l'azienda e a quella invece contenente gli impiegati ancora al suo interno.

2.1.2 Identification of the best value of k

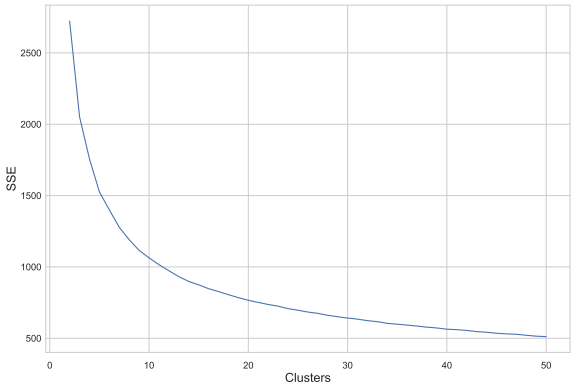
eliminare formula SSE, semplicemente SSE, al max va bene lasciare la riga di spiegazione di cos'è

Al fine di identificare il miglior numero k di clusters da utilizzare, abbiamo tenuto conto dell' *Error Sum of Squares* (SSE), ossia della somma, elevata al quadrato, della distanza tra ogni singolo data object e il centroide più vicino, ottenuta mediante la formula

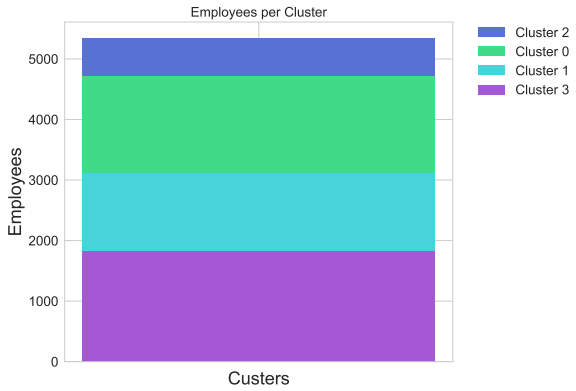
$$SSE\ =\ \sum_{i=1}^{K}\sum_{x\in C_i} dist(c_i,\ x)^2\quad ,$$

dove *dist* rappresenta la distanza Euclidea tra il centroide c_i ed il data object x. A partire da un valore iniziale di k pari a 2 fino ad un valore massimo di 50 abbiamo calcolato l'SSE risultante dall'applicazione dell'algoritmo come possiamo osservare in Figura 2.1a, e abbiamo infine deciso per un valore di k pari a 4 per l'applicazione di K-means sul data set totale, in quanto ritenuto il valore più efficiente ai fini della nostra analisi. Tale valore è stato poi confermato come il più elevato tra quelli osservati tramite lo studio del *Silhouette score*.

Perchè abbiamo deciso 4 dal grafico? Commentare/spiegare. Quanto vale la silhouette score??? Le spiegazioni non devono andare a fiducia, dobbiamo mostrare i dati risultanti e commentarli oggettivamente. Il grafico lo farei su una scala inferiore: visto che siamo interessati a un numero basso di cluster, metterei ad esempio un numero max di cluster=20 o altro valore in modo da vedere bene le variazioni nella nostra zona di interesse. (Poi nella spiegazione puoi dire che abbiamo fatto fino a 50, insomma il grafico deve essere zoommato) Sennò puoi metterne due affiancati, uno con l'andamento generale ed uno zoommato.



(a) Sviluppo dell'SSE in base all'aumentare del numero di clusters nell'applicazione dell'algoritmo K-means.



(b) Distribuzione del numero di impiegati per ognuno dei cluster scoperti durante l'analisi, in ordine decrescente per densità di popolazione.

2.1.3 Characterization of the obtained clusters

In quest'ultima sezione relativa all'algoritmo K-means descriviamo i clusters emersi durante l'analisi. Utilizzando i parametri descritti nelle sezioni precedenti, abbiamo ottenuto i clusters raffigurati in Figura 2.1b, dove possiamo osservare la densità di popolazione per ognuno dei cluster ottenuti. In Tabella 2.1 abbiamo riportato i dati caratteristici di ognuno dei cluster scoperti. Il primo cluster emerso, Cluster 0 contraddistingue gli impiegati con un alto score nelle variabili Last Evaluation e Satisfaction Level. Possiamo pensare a un tale cluster come a un gruppo di impiegati molto soddisfatti e valutati positivamente, da poco assunti, visto lo score basso in Time Spend Company. Il secondo Cluster, Cluster 1, presenta un gruppo di impiegati con uno score basso nelle variabili Average Montly Hours e Satisfaction Level. Tale cluster, essendo il terzo cluster per densità di popolazione, denota un risultato interessante e sicuramente da tenere d'occhio per i futuri sviluppi dell'azienda. Il Cluster 2 presenta un gruppo di impiegati soddisfatti e produttivi, con carichi di lavoro abbastanza elevati. L'ultimo cluster, il 3, presenta una situazione preoccupante, visto gli score molto elevati degli impiegati nelle variabili Average Montly Hours, Last Evaluation e Number Project, e il basso score nella variabile Satisfaction Level. Vedendo simili risultati è facile immaginarsi un gruppo di impiegati da poco assunti, caratterizzati da un alto tasso di ore di lavoro, il quale ha comportato il calo nel livello di soddisfazione, nonostante l'ultima valutazione fosse stata più che positiva.

Per la caratterizzazione dei clusters serve un qualche grafico che esprima quello che hai detto a parole nei commenti in modo immediato. Le tabelle sono utili come riferimento, ma andarsi a leggere sulle tabelle cosa sta succedendo non è nè immediato nè piacevole. (i valori devono essere arrotondati). Una possibilità per fare una rappresentazione sintetica è quella di fare un grafico a coordinate parallele, con sulle x le labels dei cluster e sulle y i relativi valori delle variabili, una linea per ciascuna variabile. Non so se l'ho spiegato bene, se vuoi posso provare a farlo, dimmi te. Sennò altre idee?

	Average_Montly_Hours				
Cluster	count	mean	std	min	max
0	4718.0	0.335402713014	0.119339203198	0.0	0.57
1	3112.0	0.266349614396	0.131965498666	0.0	0.88
2	5343.0	0.687986150103	0.0978323000066	0.32	1.0
3	1826.0	0.698707557503	0.218737034635	0.01	1.0

(a)

	Last_Evaluation				
Cluster	count	mean	std	min	max
0	4718.0	0.72940652819	0.154968000745	0.36	1.0
1	3112.0	0.556725578406	0.117328197911	0.36	1.0
2	5343.0	0.769792251544	0.159907000653	0.36	1.0
3	1826.0	0.796243154436	0.147510906101	0.36	1.0

(b)

	Number_Project				
Cluster	count	mean	std	min	max
0	4718.0	0.373251377702	0.171899268358	0.0	1.0
1	3112.0	0.087146529563	0.133356468399	0.0	0.8
2	5343.0	0.382481751825	0.177865527878	0.0	0.8
3	1826.0	0.730010952903	0.185110703318	0.0	1.0

(c)

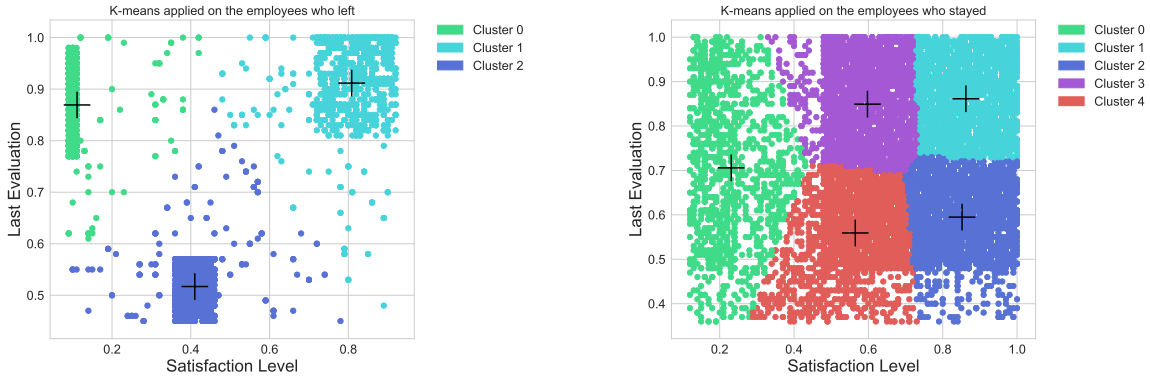
	Satisfaction_Level				
Cluster	count	mean	std	min	max
0	4718.0	0.753912674862	0.149069029588	0.29	1.0
1	3112.0	0.426690231362	0.113426667634	0.09	0.96
2	5343.0	0.746524424481	0.146280707568	0.25	1.0
3	1826.0	0.174364731654	0.103883794638	0.09	0.63

(d)

	Time_Spend_Company				
Cluster	count	mean	std	min	max
0	4718.0	0.135247986435	0.151055406452	0.0	1.0
1	3112.0	0.165690874036	0.152454170381	0.0	1.0
2	5343.0	0.216769605091	0.212362864942	0.0	1.0
3	1826.0	0.294408543264	0.14565978606	0.0	1.0

(e)

Tabella 2.1: Statistica descrittiva relativa ad ognuno dei cluster scoperti. Per ogni cluster vengono riportate le informazioni relative alla densità di popolazione, alla media, alla deviazione standard e ai valori minimi e massimi delle variabili utilizzate.



(a) (b)

Figura 2.2: Visualizzazione relativa all'applicazione dell'algoritmo K-means sul data set diviso in funzione della variabile *Left*. In Figura 2.2a è possibile osservare il clustering relativo agli impiegati che hanno lasciato l'azienda, mentre in Figura 2.2b troviamo il clustering relativo agli impiegati che sono rimasti. L'analisi dell'SSE e dello score della silhouette ha rivelato che, applicando l'algoritmo soltanto sulle variabili Satisfaction Level e Last Evaluation, il numero ideale di clusters è 3 per gli impiegati che hanno lasciato l'azienda, e 5 per gli altri.

Come ulteriore esempio, in Figura 2.2 forniamo le visualizzazioni relative all'applicazione di K-means, utilizzando le variabili Satisfaction Level e Last Evaluation, al data set diviso in base alla variabile Left. Similmente a quanto fatto per l'algoritmo applicato all'intero data set, abbiamo prima studiato l'SSE, e confrontato le nostre ipotesi con lo score fornito dall'analisi della silhouette. Come possiamo vedere nella Figura 2.2a, i 3 clusters emersi per gli impiegati che hanno lasciato l'azienda delineano un gruppo di impiegati con un basso score sia in Satisfaction Level che in Last Evaluation, un gruppo con un alto score in Last Evaluation e un basso score in Satisfaction Level e un gruppo con alto score in entrambe le variabili. Per gli impiegati ancora all'interno dell'azienda, possiamo notare nella Figura 2.2b che la situazione è decisamente più distribuita.

2.2 Hierarchical clustering

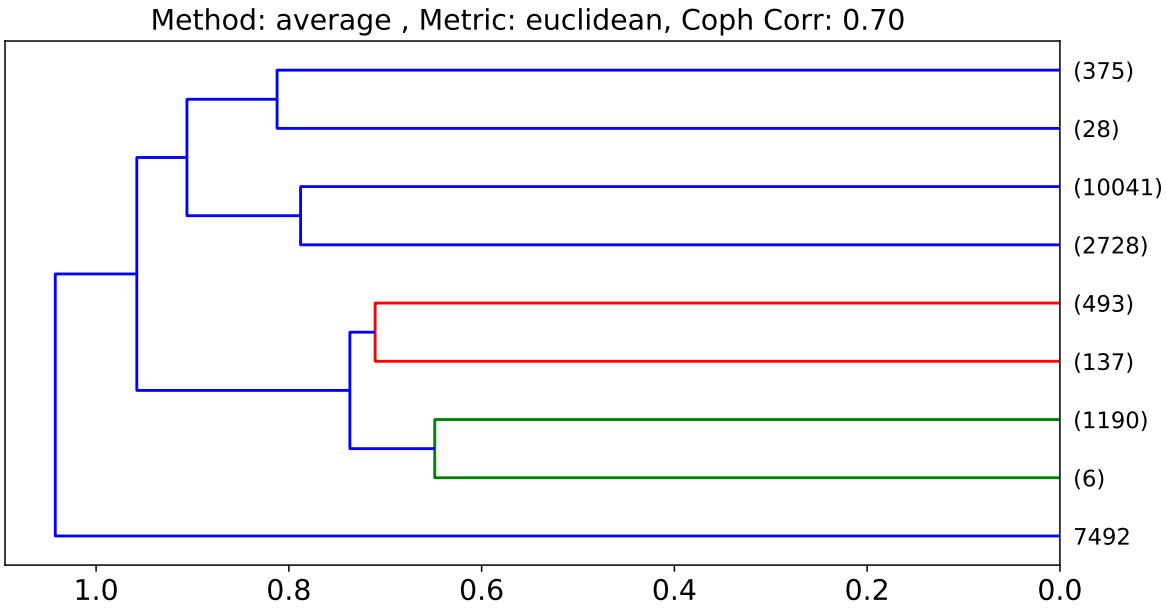


Figura 2.3: Dendrogramma per method X e metrica Y

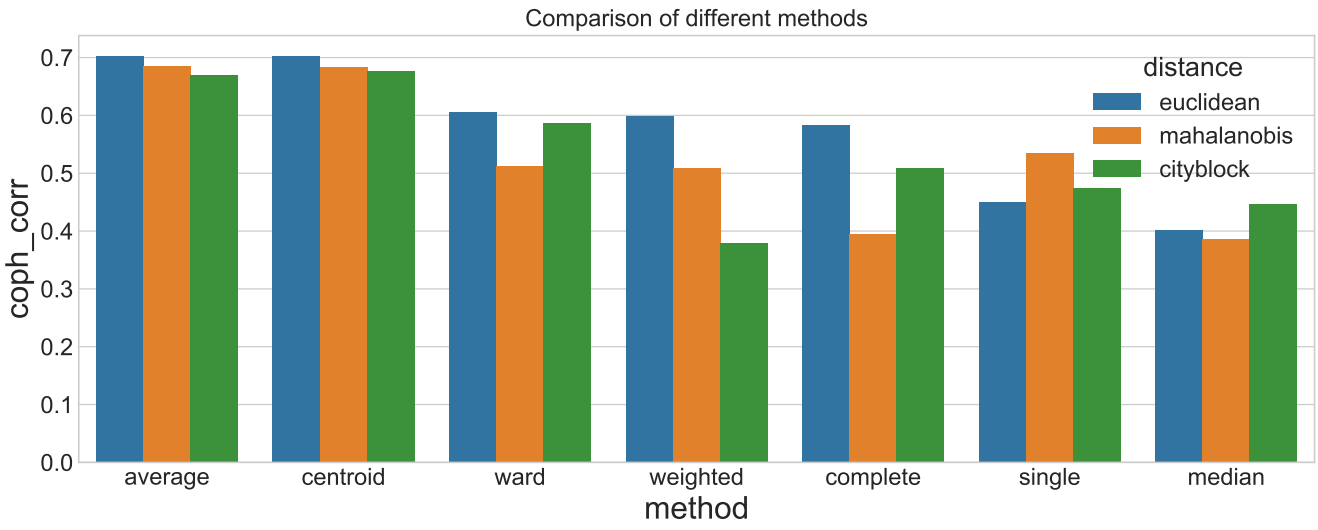


Figura 2.4: Confronto tra diversi metodi

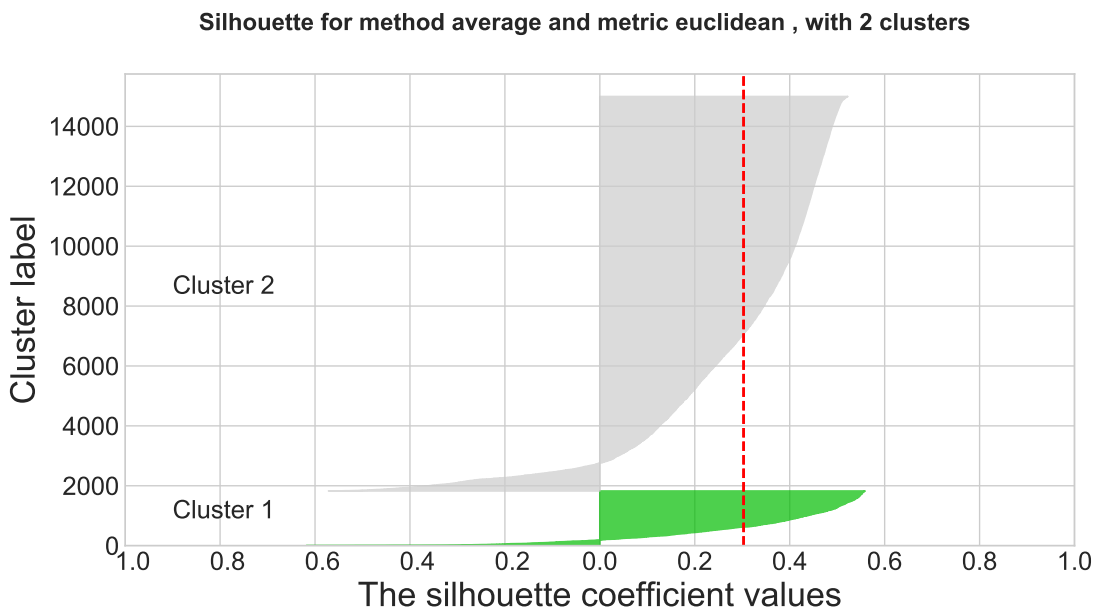


Figura 2.5: Confronto tra silhouette medie, per due clusters

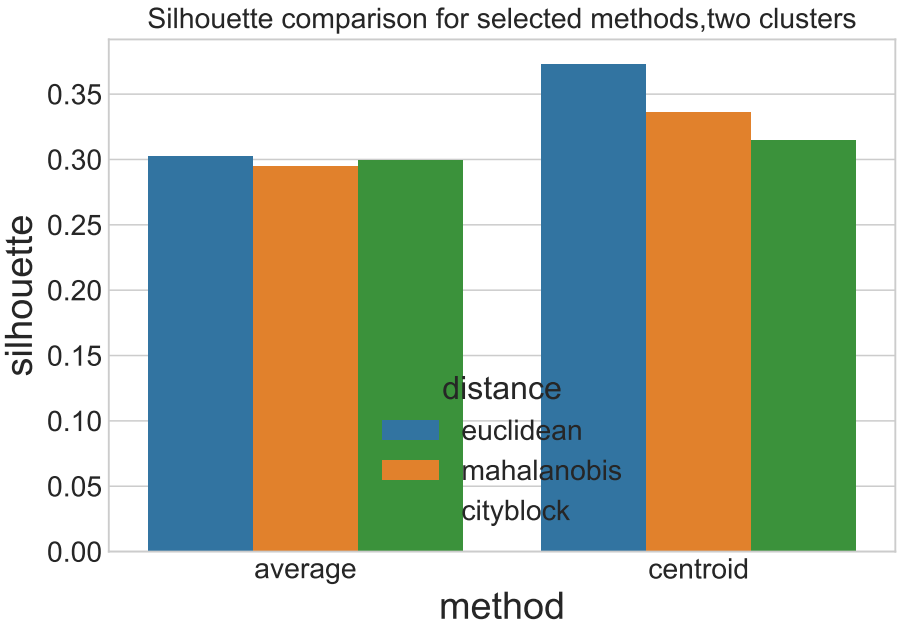


Figura 2.6: Confronto tra silhouette medie, per due clusters

3 | Association Rules

4 Classification