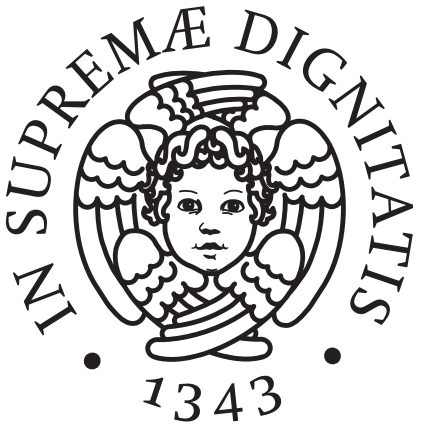


Università di Pisa

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.

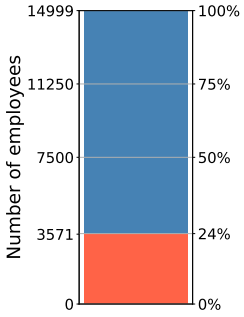


Figura 1.1: Numero di lavoratori

1.2 Data semantics

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 Kaggle [1] nella quale il dataset è pubblicato.

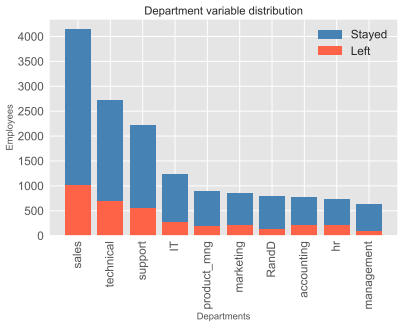
Variable	Type
Left	categorical
Department	categorical
Promotion_last_5years	categorical
Work_accident	categorical
Salary	ordinal
Satisfaction_level	continuous
Last_evaluation	continuous
Average_montly_hours	discrete
Time_spend_company	discrete
Number_projects	discrete

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 abbiamo studiato la distribuzione dei dipendenti rispetto alle variabili categoriche escludendo la distribuzione dei dipendenti rispetto a left in quanto già esplicata in precedenza nella Sezione 1.1. In Figura 1.3a 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). In Figura 1.3b, invece, rapportiamo ciascun dipendente al fatto che questo sia stato promosso negli ultimi 5 anni oppure no. Possiamo ricavare un'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% ossia circa il 0,13% dei dipendenti totali), praticamente impercettibili alla vista nel grafico. Degli impiegati rimasti, in 300 hanno ottenuto una promozione su 11428 (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. Cominciamo dalla distribuzione del salario, rappresentata in Figura 1.3c. Le percentuali che vengono indicate in Tabella 1.3d non sono in base alla totalità dei dipendenti ma riguardano solo il tipo di dipendente definito dalla colonna di appartenenza. In Figura 1.2c troviamo il rapporto tra dipendenti ancora in azienda e non, e il numero delle

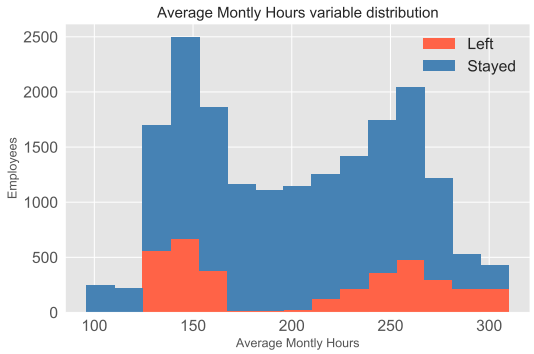
[1] https://www.kaggle.com/



(a)

Department	Stayed	Left	TotDep
sales	3126	1014	4140
technical	2023	697	2720
support	1674	555	2229
IT	954	273	1227
product_mng	704	198	902
marketing	655	203	858
RandD	666	121	787
accounting	563	204	767
hr	524	215	739
management	539	91	630
Totale	11428	3571	14999

(b)



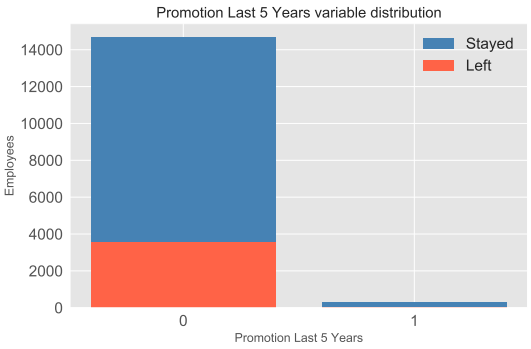
(c)

AverageMH	Stayed	Left
Media	199.06	207.42
Dev.std.	45.68	61.20
min,max	96 , 287	126 , 310

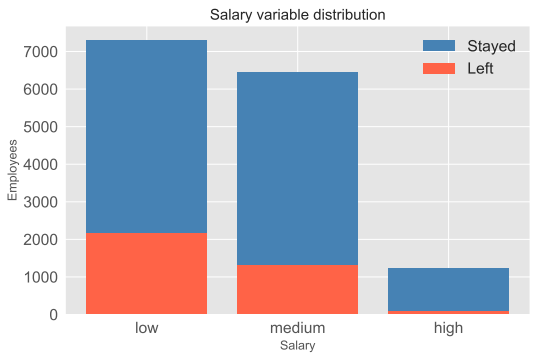
(d)

Figura 1.2: Distribuzione relativa alla variabile *Departments* (1.2a), e relativa distribuzione dei dipendenti per ogni dipartimento (1.2b), distribuzione relativa alla variabile *Average Montly Hours* (1.2c) e relativa tabella (1.2d).

(a)



(b)



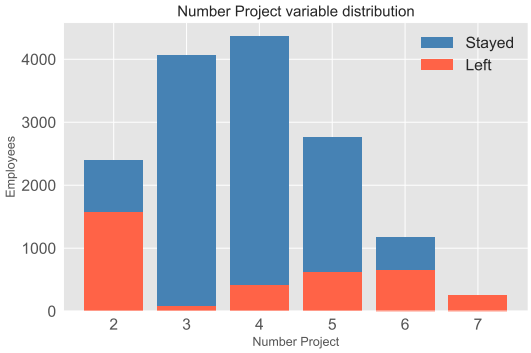
(c)

Salary	Stayed	Left
Low	5144 (∼45.02%)	2172 (∼60.8%)
Medium	5129 (∼44.88%)	1317 (∼36.9%)
High	1155 (∼10.1%)	82 (∼ 2.3%)

(d)

Figura 1.3: Distribuzioni relative alla variabile *Work Accident* (1.3a), alla variabile *Promotion Last 5 Years* (1.3b), alla variabile *Salary* (1.3c) e alla relativa distribuzione del salario per dipendente (1.3d).

ore di lavoro in media. Analizzando Figura 1.4a, l'informazione chiave che risulta da questa distribuzione è che la totalità dei dipendenti che hanno fatto 7 progetti hanno lasciato l'azienda, questo è quindi sicuramente uno dei fattori per cui i dipendenti potrebbero lasciare l'azienda. L'altro valore che risalta è i dipendenti che hanno fatto solo 2 progetti, in numero di 1567, ovvero quasi il 44% di quelli che hanno lasciato l'azienda. Di questi dovremo capire quali motivazioni li hanno portati a lasciare l'azienda, se il poco carico di lavoro o altre circostanze lavorative. Proseguiamo poi mostrando il tempo



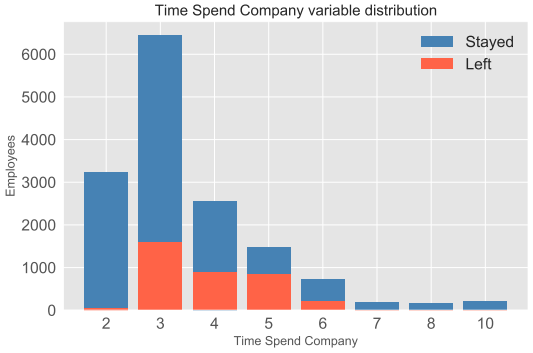
(a)

NumProject	Stayed	Left
2	821	1567
3	3983	72
4	3956	409
5	2149	612
6	519	655
7	0	256

(b)

Figura 1.4: Distribuzione relativa alla variabile *Number Project* (1.4a), e relative informazioni riguardo al numero di dipendenti per numero di progetti (1.4b).

di impiego di lavoro nell'azienda, in Figura 1.5a. Si può rilevare un fattore importante, dal settimo anno in azienda non

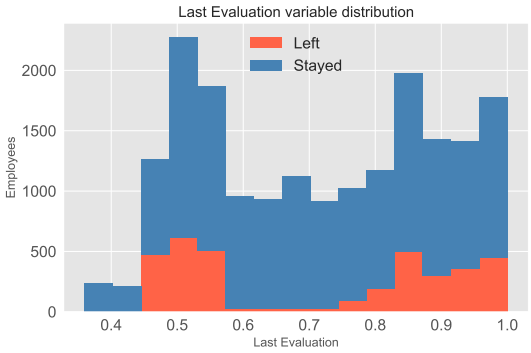


(a)

TimeSpendCompany	Stayed	Left
2	3191	53
3	4857	1587
4	1667	890
5	640	833
6	509	209
7	188	0
8	162	0
10	214	0

(b)

(c)



(d)

Satisfaction	Stayed	Left
Media	0.67	0.44
Dev.std.	0.22	0.26
min,max	0.12 , 1.0	0.09 , 0.92

(e)

LastEval	Stayed	Left
Media	0.71	0.71
Dev.std.	0.16	0.19
min,max	0.36 , 1.0	0.45 , 1.0

(f)

Figura 1.5: Distribuzione relativa alla variabile *Time Spend Company* (1.5a), e relative informazioni sui dipendenti (1.5b), distribuzioni relative alle variabili *Satisfaction Level* (1.5c) e *Last Evaluation* (1.5d), e relative tabelle (1.5e e 1.5f).

abbiamo dipendenti che hanno lasciato l'azienda. Inoltre la maggior parte del numero di dipendenti che hanno lasciato

l'azienda lo abbiamo in un range dai 3 ai 5 anni come fattore critico, con un massimo di 1587 dipendenti, ovvero quasi il 44.4% di quelli che lasciano l'azienda, nel terzo anno di lavoro. A questo punto è giusto analizzare il livello di soddisfazione dei dipendenti presente e quello della ultima valutazione, rappresentati in Figura 1.5c e Figura 1.5d.

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.

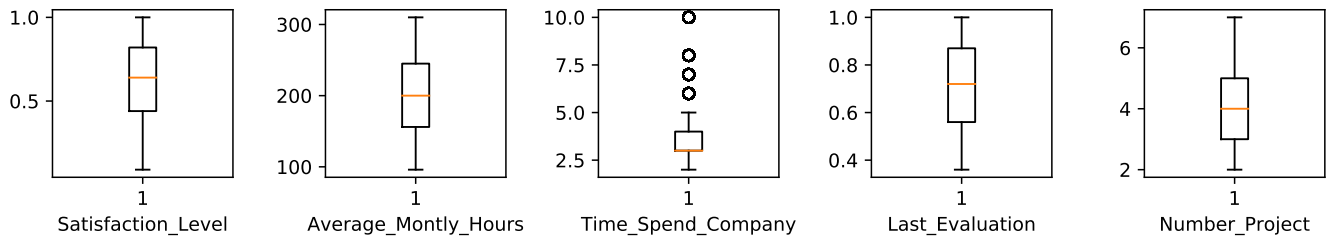


Figura 1.6: Boxplots per le variabili numeriche

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

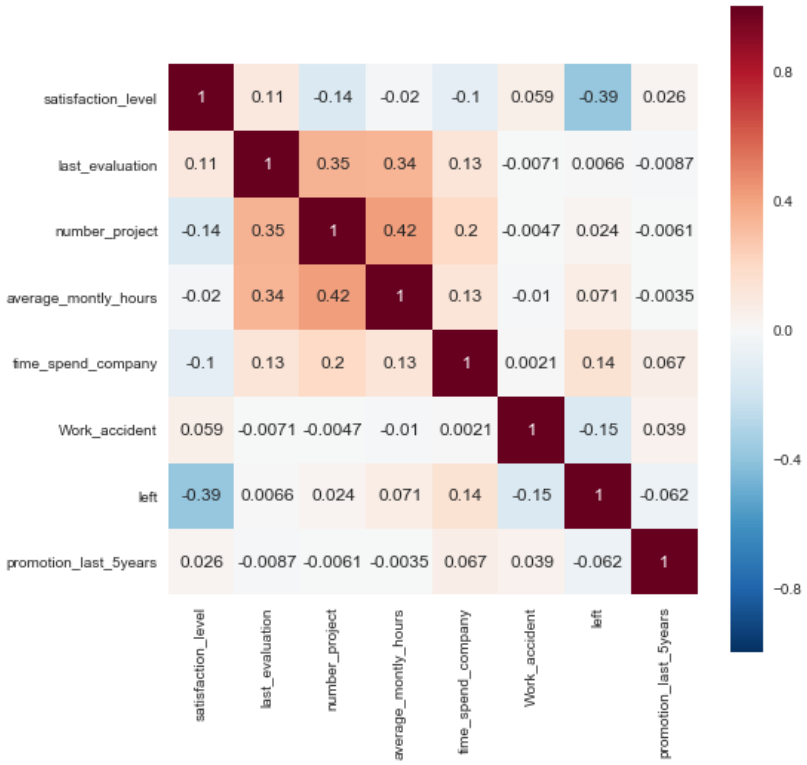


Figura 1.7: Correlation Matrix delle variabili presenti nel Dataset.

In questa sezione abbiamo studiato la correlazione ovvero la relazione lineare tra i vari attributi continuous o discreti. Dalla matrice riportata in Figura 1.7 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.

2 | Clustering analysis

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 Tabella 1.1, in modo da calcolare le distanze tra i dati in modo appropriato. 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

Come già specificato nell'introduzione a questo capitolo, abbiamo utilizzato le 5 variabili numeriche in Tabella 1.1 per il clustering. Vista la natura di tali variabili, la distance function da noi utilizzata per quantificare la distanza tra due data objects è la *distanza Euclidea*.

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), ossia della somma, elevata al quadrato, della distanza tra ogni singolo data object e il centroide più vicino. 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 2.1, dove troviamo la rappresentazione in scala ridotta a partire dal valore iniziale 2 e finale 20. 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. Il punteggio ottenuto da tale valore nello studio del *Silhouette score* è stato confrontato con gli score per gli altri valori di k, e si è rivelato essere il più alto, con un punteggio pari a 0.57.

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 2.2, 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 è formato per più di metà circa da dipendenti che hanno lasciato l'azienda e quasi metà che continuano a lavorare in questa, con un tempo di lavoro in media fra questi di poco più di tre anni. La totalità dei dipendenti che ha lasciato l'azienda (eccetto due) hanno fatto durante il periodo lavorativo esattamente due progetti. Mentre quelli rimasti hanno svolto più progetti in media e sono comunque all'interno dell'azienda da tempo ridotto, meno di tre anni. Entrambi hanno una valutazione non sufficiente.
Il secondo cluster, Cluster 1 si evince che solo 66 dipendenti su 4720 che caratterizzano questo cluster hanno lasciato l'azienda, dopo che sono rimasti a lavorare all'interno per un tempo discreto (circa 3 anni e mezzo). Il loro livello di soddisfazione è sufficiente ma nonostante abbiano un livello di valutazione elevato hanno comunque deciso di lasciare l'azienda. Mentre il livello di soddisfazione di quelli rimasti è salito. In media i dipendenti di questo cluster hanno lavorato in azienda per 3 anni.
Il terzo cluster, Cluster 2 hanno un valore bassissimo per quanto riguarda il livello di soddisfazione, si differenziano quelli che hanno lasciato l'azienda da quelli che sono rimasti per il tempo inferiore speso in azienda e il carico di lavoro più elevato, in media hanno svolto 6 progetti, in precedenza nella sezione della distribuzione abbiamo ricavato una importante informazione, che la totalità dei dipendenti che hanno svolto 7 progetti.
L'ultimo cluster, Cluster 3 è caratterizzato da un alto valore di soddisfazione, ma nonostante ciò e la valutazione sia quasi ottima, in 975 dipendenti su 5349 e che hanno speso un tempo elevato in azienda, rispetto alla media totale, decidono di lasciare l'azienda.

Da questa analisi si può evincere che:
Prima di tutto i cluster trovati fanno emergere subito che in questa azienda c'è un continuo flusso di dipendenti che entrano ed escono dalla azienda in quanto non si distinguono cluster con dipendenti che lavorano in azienda da tempo elevato. Inoltre si possono fare le seguenti supposizioni: i dipendenti che se ne sono andati nel primo cluster è perché probabilmente l'azienda non ha posto fiducia o ha dato stimoli al dipendente in modo tale che questo crescesse nell'azienda dato dal livello basso di soddisfazione.
Il dipendente lascia quasi sicuramente l'azienda quando il carico di lavoro che compie all'interno dell'azienda è elevato e questo ha un livello di soddisfazione basso, che potrebbe essere causato da una mancata promozione.

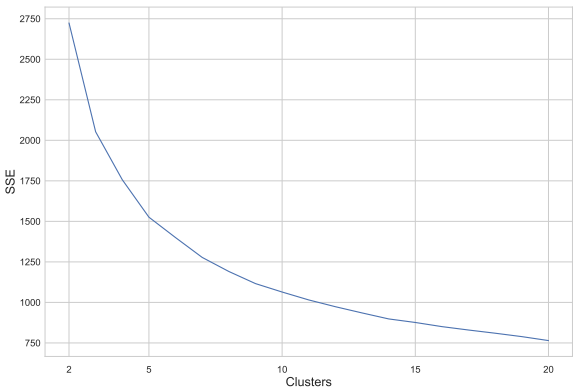


Figura 2.1

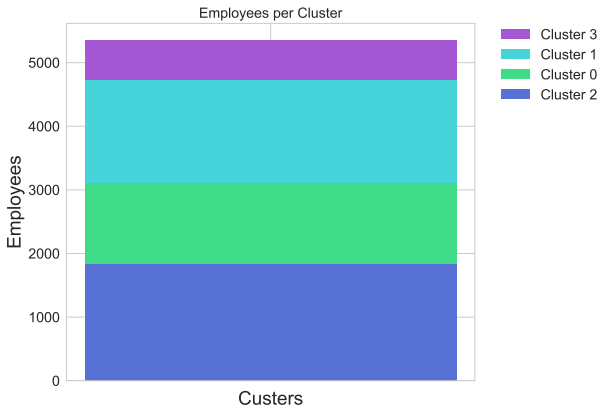


Figura 2.2

Nella Figura 2.1 viene descritto lo sviluppo dell'SSE in base all'aumentare del numero di clusters nell'applicazione dell'algoritmo K-means. Nella figura 2.2 invece la distribuzione del numero di impiegati per ognuno dei cluster scoperti durante l'analisi, in ordine decrescente per densità di popolazione.

	Average_Montly_Hours					
Cluster	countTot	meanTot	countLeft	meanLeft	countStayed	meanStayed
0	3103.0	0.26	1569.0	0.22	1534.0	0.30
1	4720.0	0.33	66.0	0.30	4654	0.33
2	1827.0	0.69	961.0	0.82	866	0.56
3	5349.0	0.68	975.0	0.70	4374	0.68

(a)

	Last_Evaluation					
Cluster	countTot	meanTot	countLeft	meanLeft	countStayed	meanStayed
0	3103.0	0.55	1569.0	0.51	1534.0	0.59
1	4720.0	0.72	66.0	0.78	4654	0.72
2	1827.0	0.79	961.0	0.85	866	0.72
3	5349.0	0.76	975.0	0.89	4374	0.74

(b)

	Number_Project					
Cluster	countTot	meanTot	countLeft	meanLeft	countStayed	meanStayed
0	3103.0	0.086	1569.0	0.0063	1534.0	0.16
1	4720.0	0.37	66.0	0.41	4654	0.37
2	1827.0	0.72	961.0	0.83	866	0.61
3	5349.0	0.38	975.0	0.49	4374	0.35

(c)

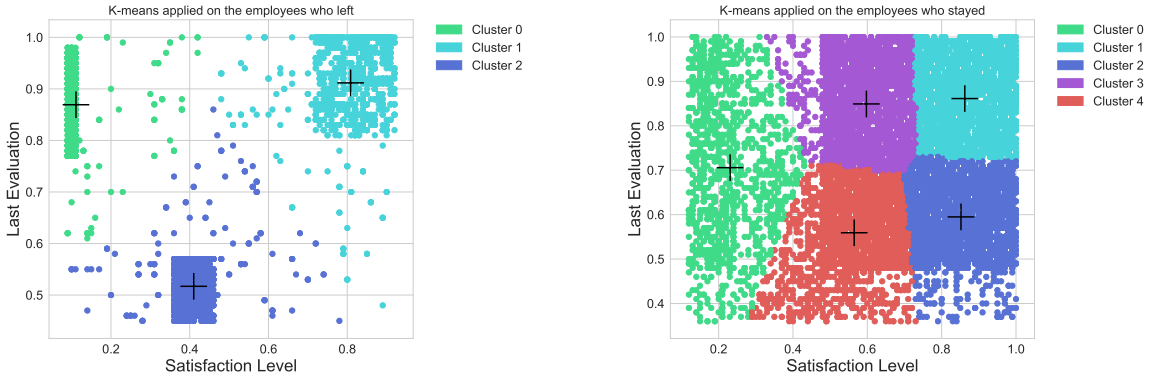
	Satisfaction_Level					
Cluster	countTot	meanTot	countLeft	meanLeft	countStayed	meanStayed
0	3103.0	0.42	1569.0	0.40	1534.0	0.44
1	4720.0	0.75	66.0	0.69	4654	0.75
2	1827.0	0.17	961.0	0.11	866	0.23
3	5349.0	0.74	975.0	0.79	4374	0.73

(d)

	Time_Spend_Company					
Cluster	countTot	meanTot	countLeft	meanLeft	countStayed	meanStayed
0	3103.0	0.16	1569.0	0.13	1534.0	0.19
1	4720.0	0.13	66.0	0.21	4654	0.13
2	1827.0	0.29	961.0	0.26	866	0.33
3	5349.0	0.21	975.0	0.38	4374	0.17

(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 2.3: Visualizzazione relativa all'applicazione dell'algoritmo K-means sul data set diviso in funzione della variabile *Left*. In Figura 2.3a è possibile osservare il clustering relativo agli impiegati che hanno lasciato l'azienda, mentre in Figura 2.3b 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 2.3 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 2.3a, 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 2.3b che la situazione è decisamente più distribuita.

2.2 Hierarchical clustering

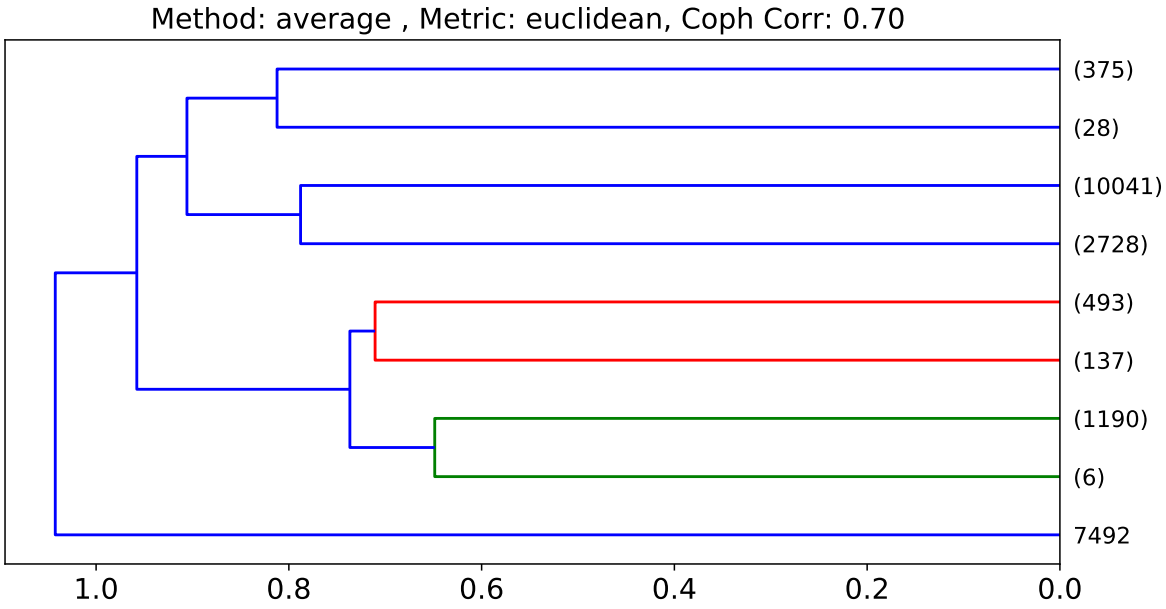


Figura 2.4: Dendrogramma per method X e metrica Y

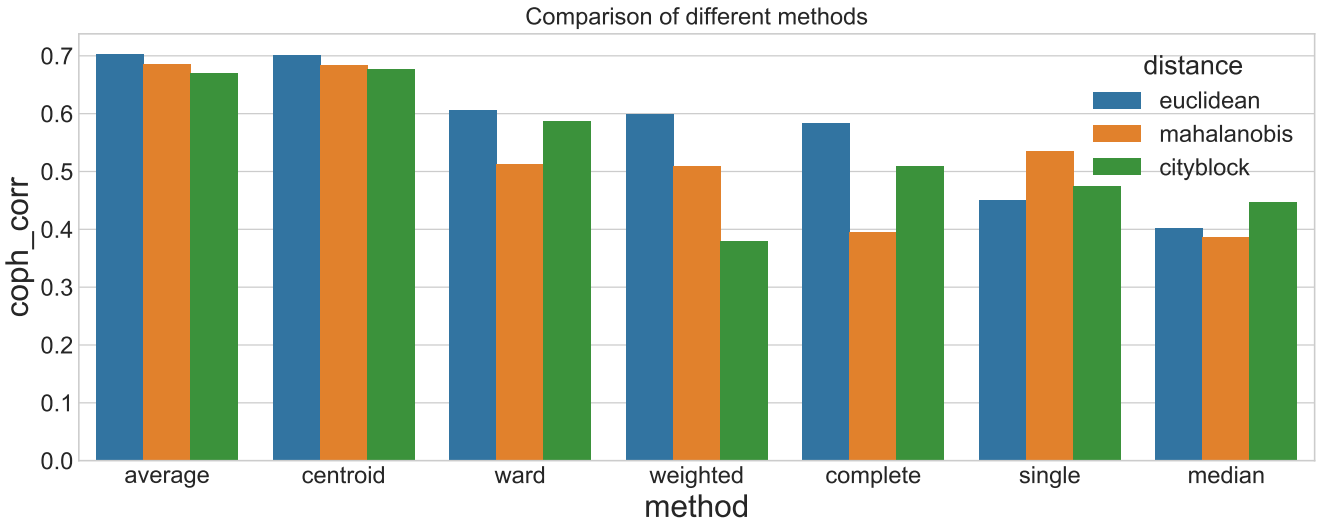


Figura 2.5: Confronto tra diversi metodi

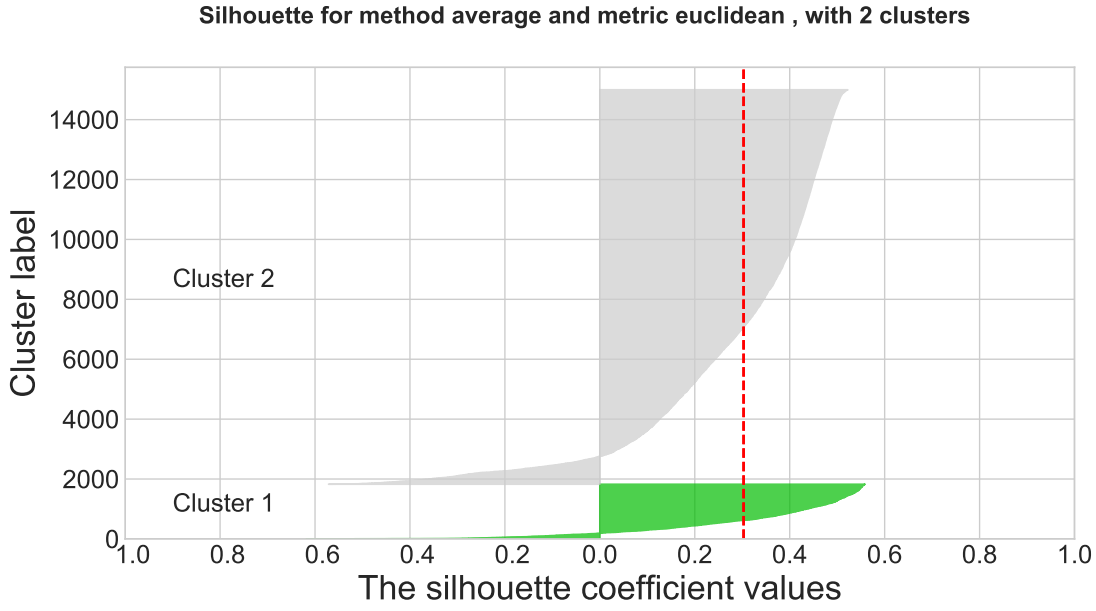


Figura 2.6: Confronto tra silhouette medie, per due clusters

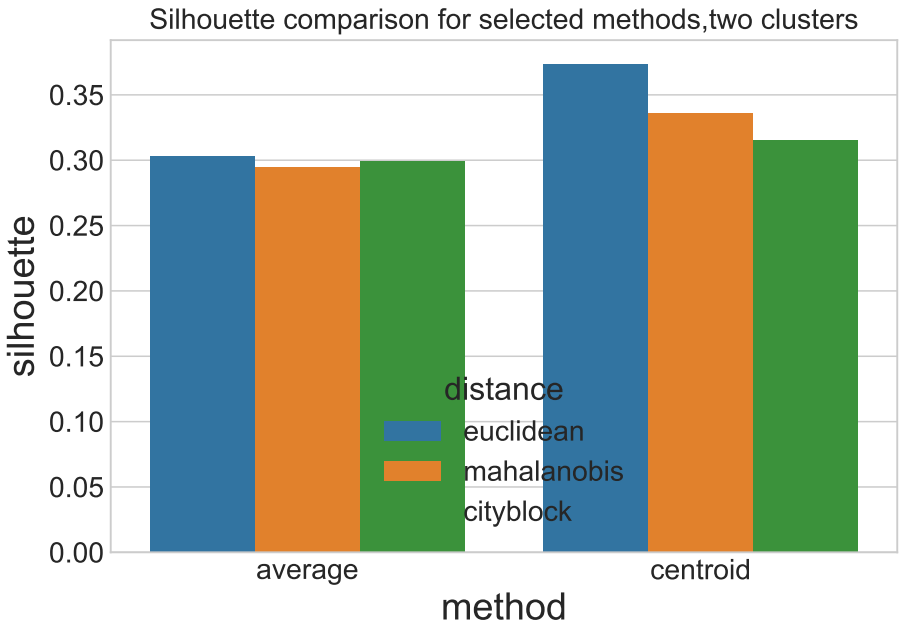


Figura 2.7: Confronto tra silhouette medie, per due clusters

3 | Association Rules Mining

Per definire le association rules prima si sono seguite le seguenti operazioni:

- Abbiamo trasformato le variabili da stringhe a numeriche, per quanto riguarda gli attributi *Salary* e *Department*.
- Abbiamo raggruppato le variabili *Last Evaluation*, *Satisfaction Level* e *Average Montly Hours*, usando 3 bins sia per la prima che per la seconda variabile, usando intervalli specifici, e 2 bins per la terza variabile, applicando anche in questo caso una suddivisione ad hoc.
- Per rendere unici i risultati numerici, è stata aggiunta una stringa subito successiva al valore numerico in modo da non renderlo ambiguo e sopratutto in modo da poter capire univocamente a che attributo si riferisca.

Aggiungere la divisione in intervalli scelta e legenda con le abbreviazioni, oppure sotto non abbreviare

3.1 Frequent patterns extraction with different values of support and different types

Dopo aver eseguito i passi preliminari descritti precedentemente abbiamo svolto l'analisi inerente ai *frequent patterns* attraverso l'applicazione dell'algoritmo *Apriori*. Per ogni iterazione dell'algoritmo, abbiamo considerato, indipendentemente dal *support*, gli itemsets con 2 o più items al loro interno. Inoltre, al fine di avere una panoramica più completa, abbiamo svolto l'analisi per i *frequent itemsets*, per i *closed frequent itemsets* e per i *maximal frequent itemsets*. Abbiamo quindi cominciato l'analisi con un support pari a 20, ossia prendendo in considerazione soltanto gli itemsets presenti in almeno il 20% della transazioni. Successivamente abbiamo utilizzato un support pari a 30. Le quantità di frequent itemsets scoperte al variare dei paramentri sono riportate in Tabella 3.1.

Support Threshold	Frequent Itemsets	Closed Frequent Itemsets	Maximal Frequent Itemsets
20	137	130	30
30	46	45	11

Tabella 3.1: Quantità di frequent itemsets trovati per ogni tipologia e support utilizzati durante l'analisi.

Come era lecito aspettarsi, esiste un rapporto di proporzionalità inversa tra la soglia di support e il numero di frequent itemsets scoperti.

3.2 Discussion of the most interesting frequent patterns

Passiamo adesso alla descrizione dei frequent items più interessanti che sono stati scoperti durante l'analisi. In Tabella 3.2 vengono riportati gli itemsets più interessanti dal punto di vista del supporto pari a 20 scoperti durante l'analisi.

Frequent Itemsets ($ST = 20$)	Support	Closed Frequent Itemsets ($ST = 20$)	Support	Maximal Frequent Itemsets ($ST = 20$)	Support
(N_WA, N_P)	0.84	(N_WA, N_P)	0.84	(standard_H, N_L, N_WA, N_P)	0.31
(N_L, N_P)	0.74	(N_L, N_P)	0.74	(intensive_H, N_L, N_WA, N_P)	0.30
(N_L, N_WA)	0.63	(N_L, N_WA)	0.63	(0_S, N_L, N_WA, N_P)	0.28
(N_L, N_WA, N_P)	0.61	(N_L, N_WA, N_P)	0.61	(1_S, N_L, N_WA, N_P)	0.28

Tabella 3.2: Frequent itemsets con supporto maggiore scoperti durante l'analisi utilizzando un supporto pari a 20. Con N_WA intendiamo l'item relativo all'assenza di incidenti sul lavoro, con N_P l'item relativo alla mancanza di promozioni, con N_L l'item relativo ai dipendenti ancora in azienda, con 0_S l'item relativo ai dipendenti con salario minimo, con 1_S l'item relativo ai dipendenti con salario medio, con intensive_H intendiamo i dipendenti con un quantitativo di ore mensili compreso tra 200 e 300 e con standard_H intendiamo i dipendenti con un quantitativo di ore mensili inferiore a 200.

Descriviamo per primi i frequent itemsets e i closed frequent itemsets, visto che sono identici. Possiamo notare come la situazione presentata proponga in maggioranza impiegati i quali non hanno subito incidenti sul lavoro, che non sono stati promossi e che non hanno lasciato l'azienda. Per quanto riguarda i maximal frequent itemsets troviamo che gli impiegati con carichi di lavoro sia standard che elevati, che non hanno lasciato l'azienda, non hanno avuto incidenti sul lavoro e che non sono stati promossi negli ultimi 5 anni sono i più diffusi, seguiti dagli impiegati di salario minimo e medio, non promossi e i quali non hanno avuto incidenti sul lavoro. Portando la soglia del support a 30, gli itemsets più diffusi sono gli stessi che sono stati descritti per la soglia pari a 20, evitiamo quindi di descriverli.

3.3 Association rules extraction with different values of confidence

Per estrarre le associazioni interessanti del dataset sono stati utilizzati due approcci principali per la scelta dei parametri dell'algoritmo *apriori*, in modo da ottenere:

- *Regole "generali"*, ovvero associazioni interessanti valide per un numero ampio di impiegati, ottenute fissando un alto supporto minimo, pari a $MinSupp = 20\%$, nella ricerca degli itemset frequenti.

- *Regole "specifiche"*: il supporto minimo è stato fissato considerando che uno degli obiettivi principali delle analisi contenute in questo report è capire il perchè una parte consistente dei dipendenti ha lasciato l'azienda. La percentuale di dipendenti che hanno lasciato l'azienda corrisponde al 24% del totale, assumendo come significativa una regola che riguardi almeno il 20% dei dipendenti che hanno lasciato l'azienda, risulta un supporto minimo pari a circa $MinSupp = 5\%$.

L'analisi è stato eseguita per entrambi i valori di $MinSupp$ indicati, variando la confidenza minima *MinConf* per valori compresi tra $50 - 100\%$. In Fig. sono riportate il numero di regole ottenute, in funzione della confidenza. Le regole cercate hanno un solo item come parte conseguente, per facilitare l'analisi e ridurre il numero di regole in partenza. Per ciascuna regola estratta sono stati calcolati gli indici di *confidence* e *lift* in modo da valutare l'interesse oggettivo delle associazioni trovate.

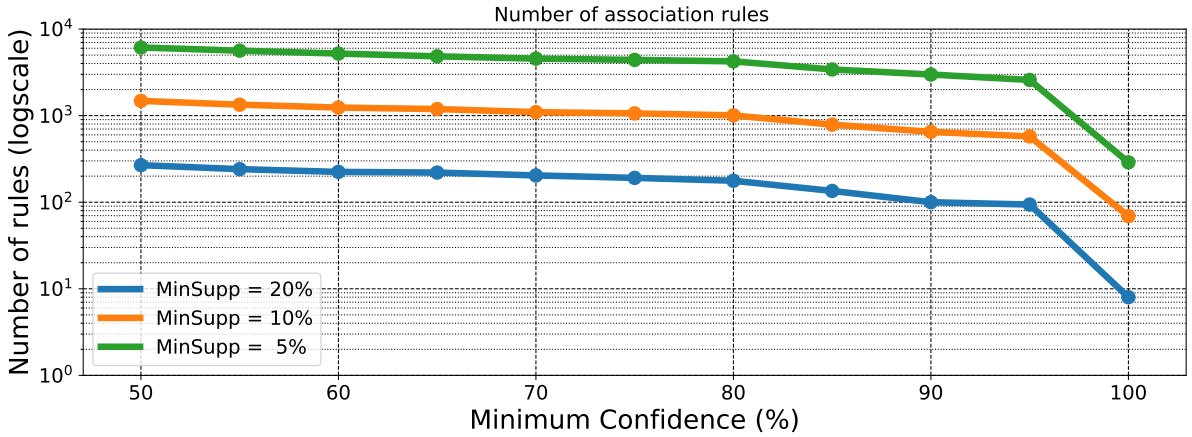


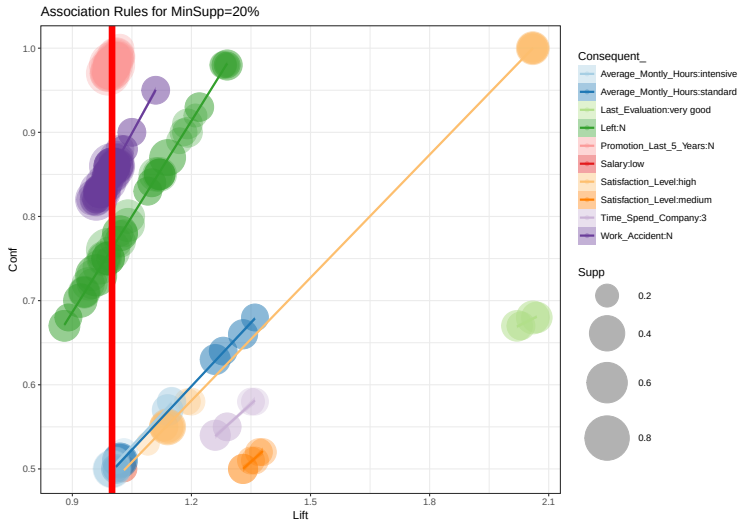
Figura 3.1: Numero di regole estratte al variare di *MinConf*, per differenti valori di *MinSupp*.

3.3.1 Regole generali

Le regole generali sono state cercate fissando a priori *Minsupp=20%* e considerando un valore di confidenza minimo *MinConf=50%* , scelto osservando l'andamento in Fig. 3.1, sufficiente per ottenere un numero trattabile di regole.

Per un valore di *MinConf=50%* sono state ottenute 268 regole rappresentate in Fig. 3.2, con i rispettivi valori di Lift, Confidence e Support. Si osserva che la maggior parte delle regole ha un Lift basso, minore di 1.5. Le regole sono state rappresentate raffigurando il valore di $Conf$ in funzione di $Lift$. Questo permette di identificare visivamente in modo immediato le regole con la stessa parte conseguente, poichè in tal caso i due indici sono proporzionali con coefficiente pari al supporto della parte conseguente. Le regole che possono essere considerate oggettivamente interessanti, con $Lift > 1.5$, visibili nella parte destra del grafico sono riportate in Tab 3.3 e mostrano un legame positivo tra un alto livello di valutazione (*very good, Last Evaluation*) ed un alto grado di soddisfazione degli impiegati, a cui si accompagnano nella parte antecedente la mancanza di incidenti sul lavoro (*N WA*) ed il fatto di essere rimasti in azienda.

Tutti gli impiegati che hanno una *'Last evaluation: very good'* hanno anche un alto tasso di soddisfazione ('high_SL'), mentre il viceversa accade nel 68% dei casi. Queste regole non sono particolarmente interessanti dal punto di vista dell'analisi, poichè abbastanza ovvie, e non aggiungono molto alle analisi statistiche eseguite.



Consequent	Antecedent	Supp	Conf	Lift
high_SL	('very good_LE',)	0.329	1.000	2.061
high_SL	('very good_LE', 'N_P')	0.321	1.000	2.061
high_SL	('very good_LE', 'N_L')	0.281	1.000	2.061
high_SL	('very good_LE', 'N_WA')	0.278	1.000	2.061
high_SL	('very good_LE', 'N_L', 'N_P')	0.273	1.000	2.061
high_SL	('very good_LE', 'N_WA', 'N_P')	0.272	1.000	2.061
high_SL	('very good_LE', 'N_L', 'N_WA')	0.232	1.000	2.061
high_SL	('very good_LE', 'N_L', 'N_WA', 'N_P')	0.226	1.000	2.061
very good_LE	('high_SL', 'N_WA')	0.278	0.681	2.070
very good_LE	('high_SL', 'N_WA', 'N_P')	0.272	0.680	2.068
very good_LE	('high_SL',)	0.329	0.678	2.061
very good_LE	('high_SL', 'N_P')	0.321	0.677	2.059
very good_LE	('high_SL', 'N_L', 'N_WA')	0.232	0.668	2.030
very good_LE	('high_SL', 'N_L', 'N_WA', 'N_P')	0.226	0.667	2.026
very good_LE	('high_SL', 'N_L')	0.281	0.666	2.025
very good_LE	('high_SL', 'N_L', 'N_P')	0.273	0.665	2.023

Tabella 3.3: Association rules per $MinSupp = 20\%$ e $Lift > 1.5$

Figura 3.2: Grafico,mettere LEGENDA PER COLORE E SIZE (Support)

3.3.2 Regole specifiche

Usando $MinSupp = 5\%$ naturalmente il numero di associazioni trovate cresce notevolmente, raggiungendo in questo caso le oltre 6000 regole. Le regole sono state dunque filtrate ulteriormente in modo da avere un interesse oggettivo considerando $Lift > 1$ e $Conf > 0.9$, ed inoltre $Supp < 20\%$, poichè per supporto maggiore sono state analizzate in precedenza. Le circa 2000 regole ottenute sono rappresentate in Fig. 3.3, si può osservare che le regole con Lift più alto hanno come parte conseguente, partendo dalla destra del grafico, *Satisfaction Level: low , Number of Project: 2, Left : Y, Last Evaluation: Insufficient*.

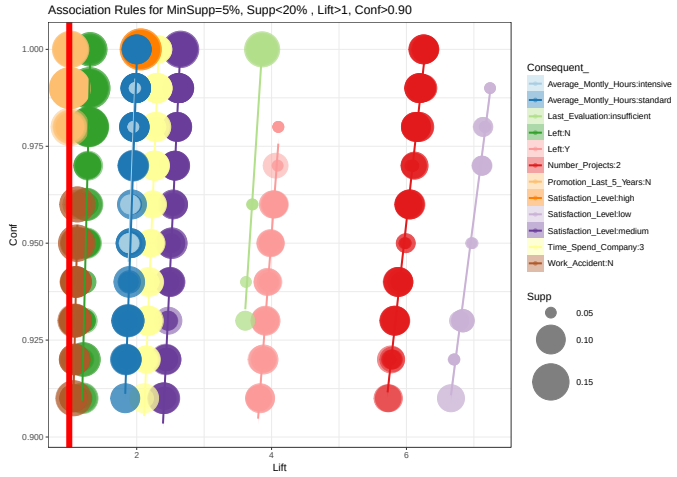


Figura 3.3: Grafico,mettere LEGENDA PER COLORE E SIZE (Support)

Regole per impiegati che hanno lasciato

Tra le regole specifiche sono state estratte solo quelle con parte conseguente uguale a *Left:YES*. Per filtrarle ulteriormente è stato usato un ulteriore criterio, che riportiamo insieme a tutti gli step successivi che hanno portato alla selezione delle regole più interessanti, riguardanti gli impiegati che hanno lasciato l'azienda, riportate in Fig. 3.4:

1. $MinSupp = 5\%$ e $MinConf = 50\%$ nell'algoritmo *apriori*, ottenendo $N_{rules} = 264$ per la parte conseguente $Left : Yes$;
2. rimozione delle regole contenenti $WorkAccident : No$, poichè non è semanticamente sensata l'implicazione di $Left : Yes$, ottenendo $N_{rules} = 136$;
3. filtro a posteriori con $Lift > 1$ e $Conf > 0.85$, ottenendo $N_{rules} = 62$

4. Per le regole con itemsets delle parti antecedenti sottoinsiemi l'uno dell'altro e *con supporto uguale* (confronto all'$1/1000$), sono state considerate le regole con parte antecedente più lunga, Ad esempio per due regole del tipo $r_1 : A \to Y$, $r2 : AB \to Y$, con A, B, Y differenti item, e $Supp_1 = Supp_2$ si seleziona solo la regola r_2.

Le 12 regole ottenute con il procedimento spiegato sono riportate in Fig. 3.4 insieme ai rispettivi valori di confidence e support. E' inoltre riportato il supporto relativo alle persone che hanno lasciato, $Supp\,Left$. Le regole sono rappresentate in verticale, con i valori 10 e 0 che indicano la presenza o meno di un item nella parte antecedente della regola.

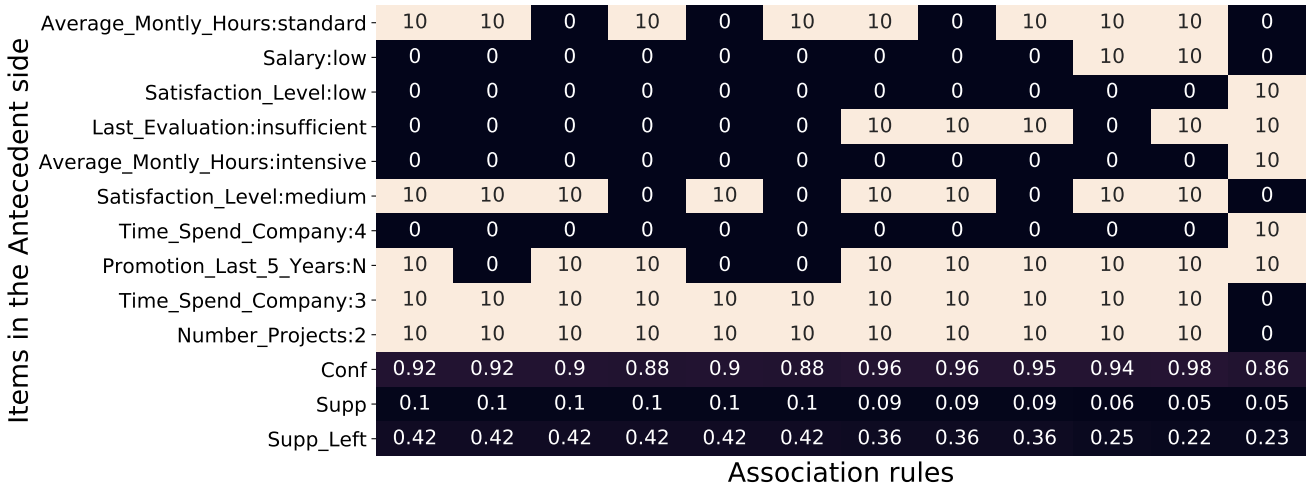


Figura 3.4: Regole associative più interessanti con parte conseguente gli impiegati che hanno lasciato l'azienda. I valori 10-0 corrispondono alla presenza o meno di un item nella parte antecedente della regola

Il supporto massimo risulta pari al 10%, corrispondente ad una percentuale di circa il 40% tra gli impiegati che hanno lasciato. Si osserva dunque una correlazione positiva tra gli item riportati in Fig. 3.4 e il fatto di aver lasciato l'azienda.

In generale le regole mostrano che le persone che lasciano l'azienda sono poco stimolate, risulta che circa il 40% degli impiegati che hanno lasciato l'azienda lavora a solo due progetti, con un impegno di ore mensili basso, una soddisfazione media, ed una presenza in azienda da 3 anni. Altri aspetti rilevanti sono la mancanza di una promozione, una valutazione insufficiente ed un salario basso.

Altre regole interessanti

Dalle regole specifiche estratte ne sono state ricavate alcune interessanti, per alto valore di $Conf$ e $Lift$, riportate in Tab. 3.4

Consequent	Antecedent	Supp	Conf	Lift
low_SL	('4_T', 'Y_L', 'insufficient_LE', 'intensive_H')	0.054	0.990	7.245
2_NP	('Y_L', 'medium_SL', '3_T', 'standard_H', 'N_P')	0.100	0.996	6.256
insufficient_LE	('low_SL', 'intensive_H', 'N_P')	0.096	1.000	3.861

Tabella 3.4: Altre regole interessanti

3.4 Use the most meaningful rules to predict if an employee will leave prematurely or not and evaluate the accuracy

Date le varie association rules trovate dalle varie prove queste sono quelle più significative per predirre se un impiegato lascerà prematuramente l'azienda oppure no:

- AR ... accuratezza trovata: - AR1 - AR2

Da queste possiamo scaturire che un impiegato lasci il posto di lavoro prematuramente quando è nelle seguenti condizioni:
Invece rimarrà quando avrà una condizione del tipo:

4 | Classification