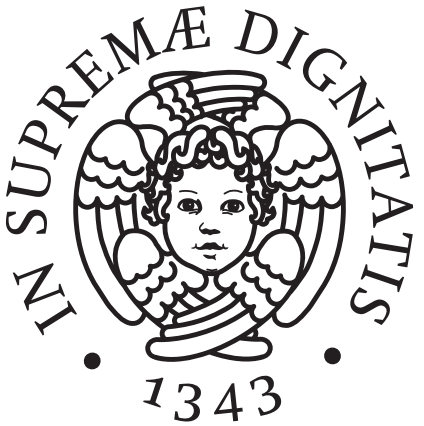


Università di Pisa

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.

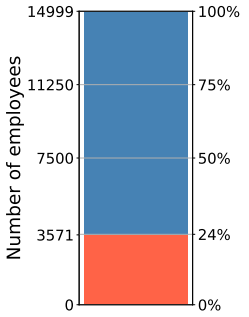


Figura 1.1: Numero di lavoratori

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 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 1.1.

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

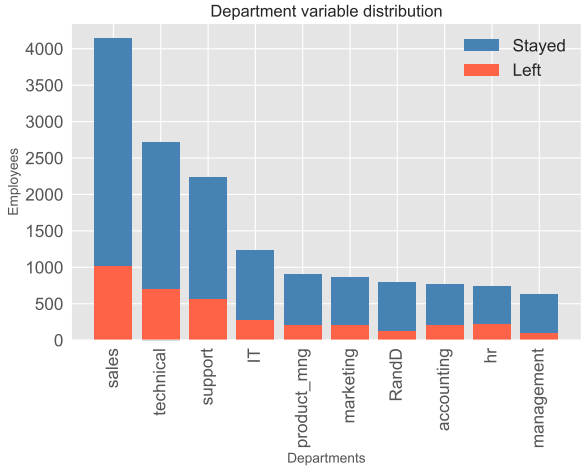


Figura 1.2: Distribuzione relativa alla variabile *Department*

Department	Stayed	Left	TotDep
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 dipendenti per *Department*.

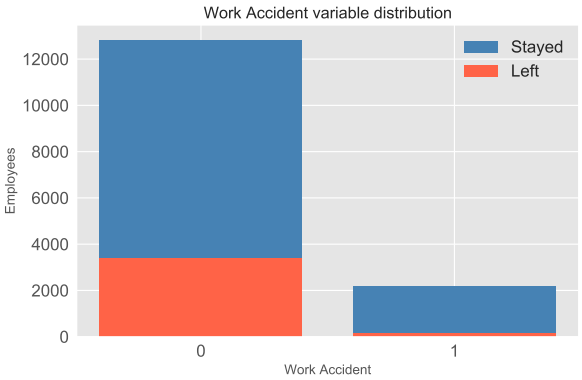


Figura 1.3: Distribuzione relativa alla variabile *Work Accident*

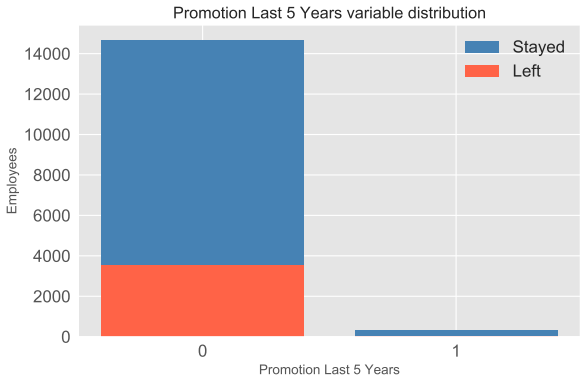


Figura 1.4: Distribuzione relativa alla variabile *Promotion Last 5 Years*

In Figura 1.3 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 1.4 invece rapportiamo 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:

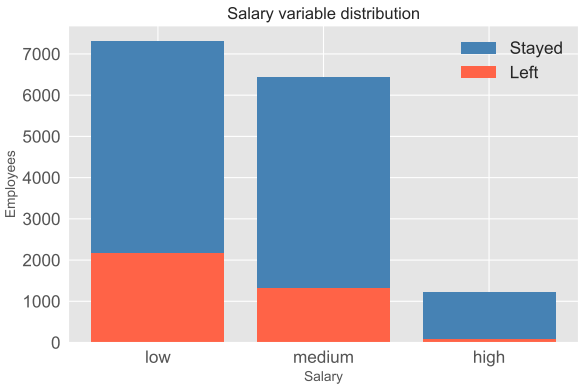


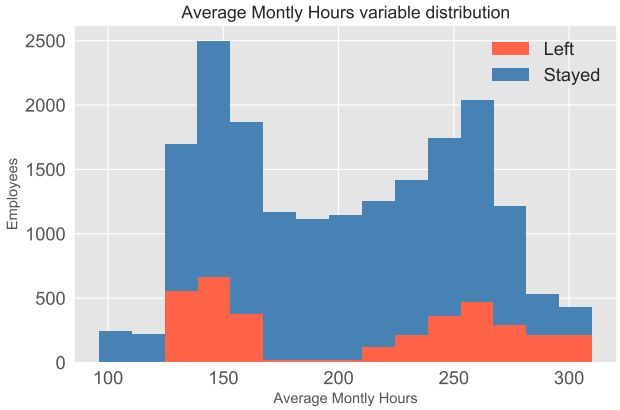
Figura 1.5: Distribuzione relativa alla variabile *Salary*

Salary	Stayed	Left
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.

Le percentuali che vengono indicate nella Tabella 1.3 non sono in base alla totalità dei dipendenti ma riguardano solo il tipo di dipendente definito dalla colonna di appartenenza.

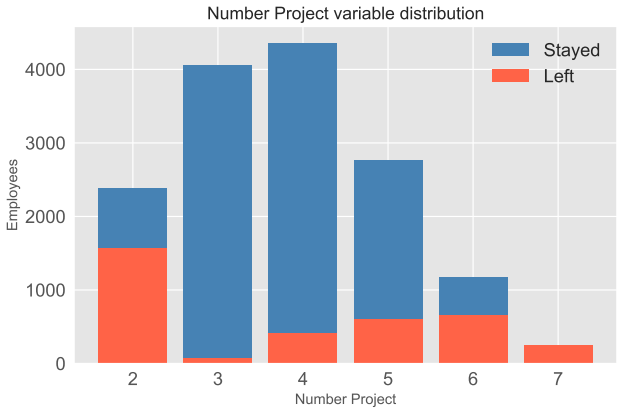
Proseguendo poi con il rapporto tra dipendenti ancora in azienda e non, e il numero delle ore di lavoro in media:



AverageMH	Stayed	Left
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 Montly Hours*

Dipendenti rimasti e non rispetto al numero di progetti:

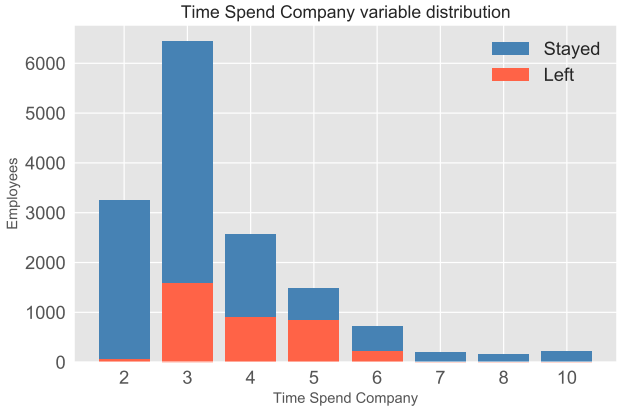


NumProject	Stayed	Left
2	821	1567
3	3983	72
4	3956	409
5	2149	612
6	519	655
7	0	256

Figura 1.7: Distribuzione relativa alla variabile *Number Project*

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.

Per proseguire con il tempo di impiego di lavoro nell'azienda:



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

Figura 1.8: Distribuzione relativa alla variabile *Time Spend Company*

Si può rilevare un fattore importante, dal settimo anno in azienda non 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:

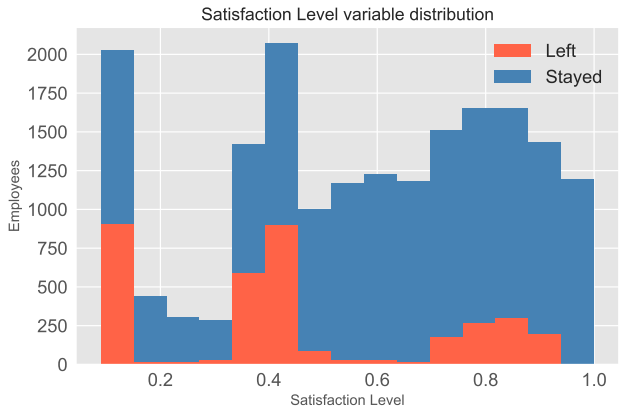


Figura 1.9: Distribuzione relativa alla variabile *Satisfaction Level*

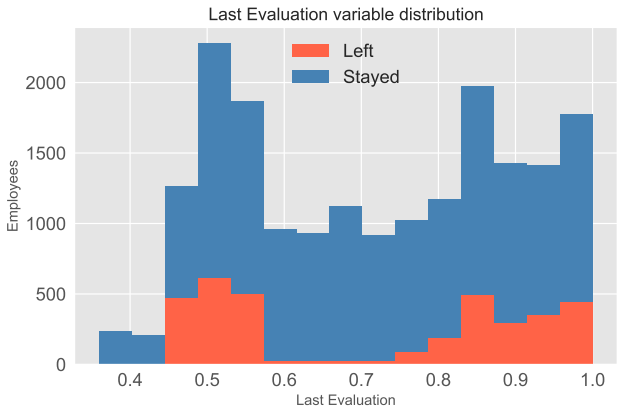


Figura 1.10: Distribuzione relativa alla variabile *Last Evaluation*

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

LastEval	Stayed	Left
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

- 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.

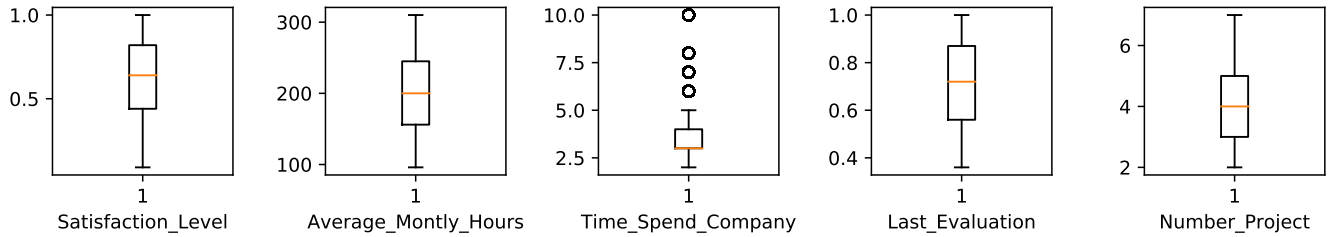


Figura 1.11: Boxplots per le variabili numeriche

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

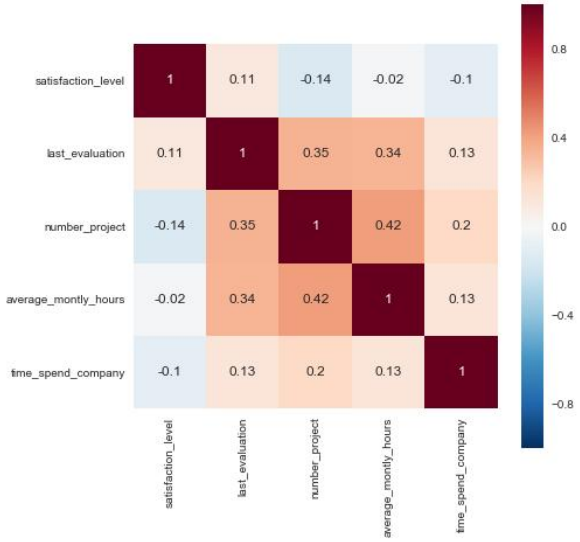


Figura 1.12: Correlation Matrix delle variabili presenti nel Dataset.

In questa sezione abbiamo studiato la correlazione ovvero la relazione lineare tra i vari attributi continuous o discreti. Dalla matrice riportata in Figura 1.12 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 last_evaluation, number_project e average_monthly_hours. Queste 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 satisfaction_level con number_project e time_spend_company.

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 2.4, 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.

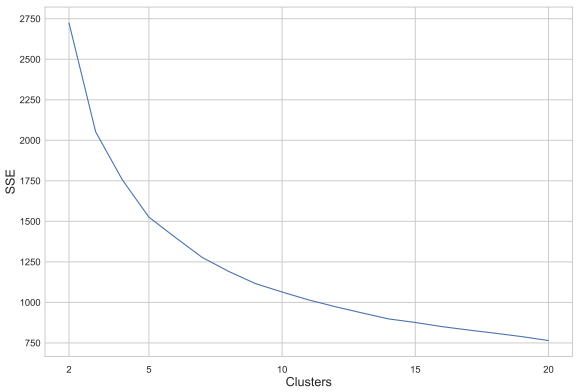


Figura 2.1

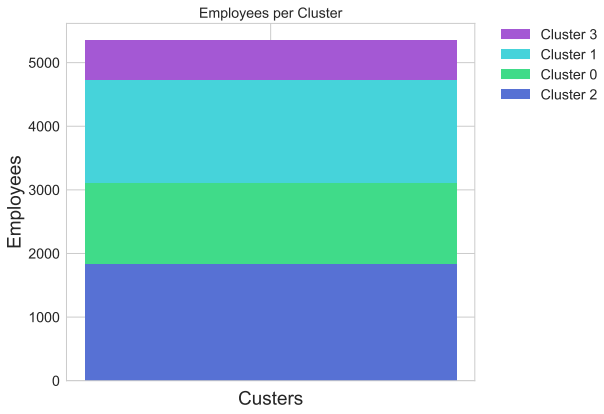


Figura 2.2

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)

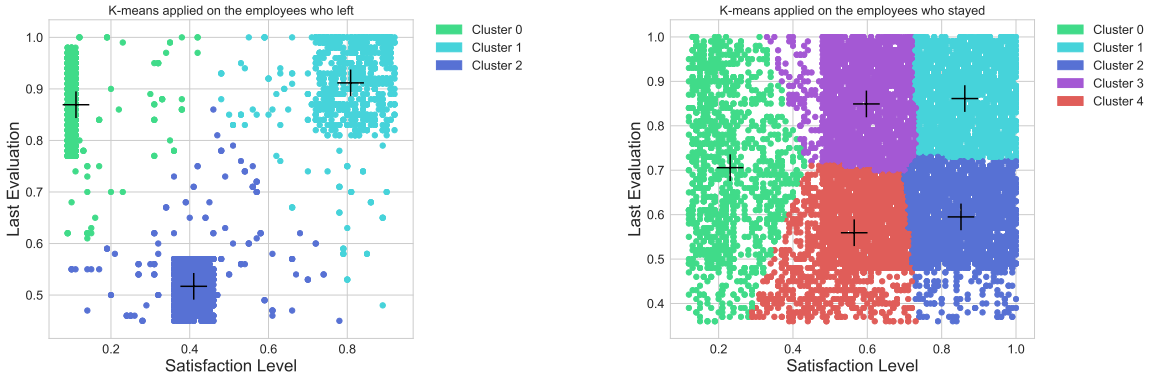
	Satisfaction_Level					
Cluster	countTot	meanTot	countLeft	meanLeft	countStayed	meanStayed
0	3103.0	0.42	1569.0	0.40	1534.0	0.44
1	4720.0	0.75	66.0	0.69	4654	0.75
2	1827.0	0.17	961.0	0.11	866	0.23
3	5349.0	0.74	975.0	0.79	4374	0.73

(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 DBSCAN

2.2.1 Choice of attributes and distance function

Il DBSCAN è un metodo di clustering basato sulla densità. Parlando di densità si intende il numero di punti in un specifico raggio chiamato eps. Per seguire la stessa analisi riportata dal Kmeans utilizziamo anche in questa metodologia tutti gli attributi eccetto quelli categorici e quello ordinale $salary$.

2.2.2 Study of the clustering parameters

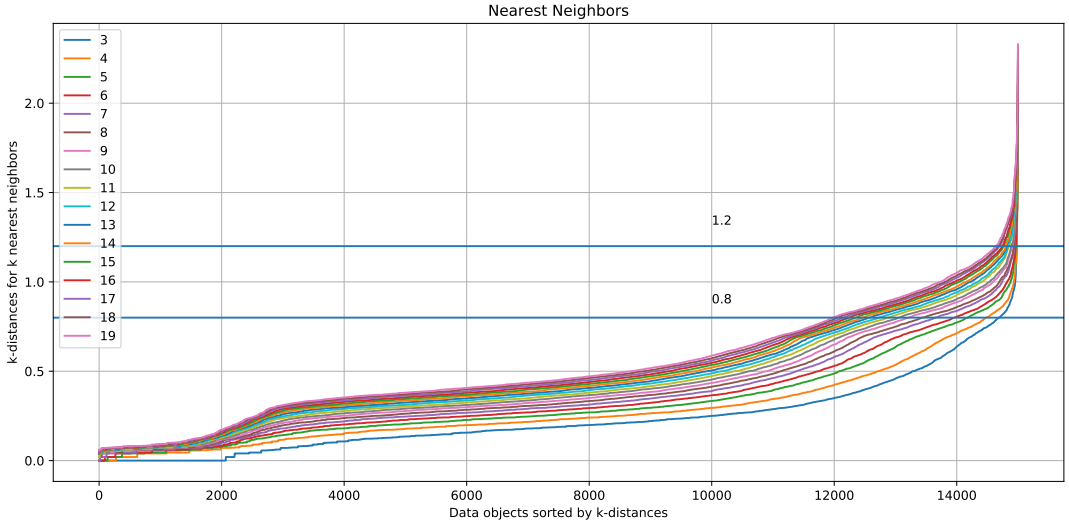


Figura 2.4: Andamento numero di cluster per ciascun eps

Per la scelta dei parametri del DBSCAN ovvero : il raggio di distanza eps dai punti centrali, chiamati centroidi, e il minimo numero di punti in questo raggio $minSamples$, si è definito l'andamento del rapporto tra eps e minSamples attraverso lo studio del k-nearest neighbors. Dal grafico abbiamo individuato il punto di flesso in un range di eps tra 0.18 e 0.25. L'analisi prendeva anche in considerazione il valore della silhouette, nell'intervallo citato precedentemente si è trovato che il flesso si trovava con $eps = 0.22$ (il valore esatto è 0.227778), $minSamples = 13$, ottenendo così 3 cluster, e un valore di $silhouette$ di 0.310. Per determinare la scelta dei parametri abbiamo deciso di utilizzare la distanza euclidea per rendere più immediato il confronto con le altre metodologie di clustering. La combinazione trovata è stata scelta dopo aver testato anche soluzioni che mostrassero più cluster rispetto a tre, ma sono state scartate in quanto queste indicano sempre la medesima situazione: un cluster principale in cui troviamo quasi la totalità dei dipendenti e negli altri cluster un numero sempre più esiguo di dipendenti, quindi non danno informazioni aggiuntive particolari rispetto a quella riportata di seguito.

2.2.3 Characterization and interpretation of the obtained clusters

Cluster	count	Satisfaction	L.Evaluation	Num Projects	Average Montly Hours	Time S. Company
0	14738.0	0.61	0.71	0.36	0.49	0.17
1	134.0	0.68	0.74	0.25	0.46	1
2	8.0	0.75	0.78	0.4	0.8	1

Tabella 2.2: Statistica descrittiva relativa ad ognuno dei cluster scoperti con la metodologia DBSCAN. Per ogni cluster vengono riportate le informazioni relative alla media delle variabili continue.

I cluster rilevati da questa metodologia sono 3: Il primo cluster, Cluster 0 è caratterizzato da 14738 dipendenti che descrivono la situazione generale dell'azienda, si nota subito dal TS, Time Spend Company, che i dipendenti in media rimangono per breve tempo in azienda. I restanti valori medi per gli altri attributi descrivono una situazione nella media sufficiente come soddisfazione, con un buon numero di progetti, una valutazione in media discreta, e un numero medio di ore di lavoro nella norma. Il secondo,il terzo mostrano caratteristiche simili: Entrambi i tipi di dipendenti rispettivamente 134 per il secondo cluster e 8 per il terzo, sono al decimo anno di lavoro nell'azienda e stanno ancora lavorando in questa, però si articolano in modo differente per gli altri attributi. Infatti nel secondo cluster, Cluster 1 raggruppa dipendenti soddisfatti sufficientemente, con un livello di valutazione discreto, e un numero di ore medie poco meno della media totale e infine un numero di progetti sufficiente. Mentre nel terzo cluster , Cluster 2 abbiamo una situazione diversa alla precedente per quanto riguarda il numero di ore di lavoro, questi sono i dipendenti che lavorano più ore in media al mese e lavorano su più progetti. I restanti119 dipendenti sono stati considerati come punti di noise.

2.3 Hierarchical clustering

Sulle variabili numeriche del dataset è stato eseguito un clustering gerarchico di tipo agglomerativo, con l'utilizzo di diverse metriche e differenti metodi per il calcolo della dissimilarità tra clusters. Il confronto tra i diversi metodi è stato effettuato utilizzando come indice di validità, la *cophenetic correlation* ρ_{coph}. Ciascuna coppia di dati del dataset è unita nel dendrogramma in un nodo ad una distanza denominata *cophenetic distance*. La cophenetic correlation è calcolata come coefficiente di correlazione di Pearson tra gli $n(n-1)/2$ valori della matrice delle distanze e le corrispondenti cophenetic distances. Il valore di tale correlazione può essere interpretato come il grado di aderenza del dendrogramma alle distanze originali, ed è quindi una misura della sua validità. In figura 2.6 si osserva che tra i metodi utilizzati *average* e *centroid* risultano quelli con il valore massimo della cophenetic correlation ed equivalenti tra loro. La dissimilarità calcolata con il metodo *average* consiste nella media di tutte le distanze tra coppie di dati appartenenti a due clusters diversi, mentre il metodo *centroid* utilizza la distanza tra i centroidi. Per ciascun metodo sono state testate tre metriche, quella euclidea, Manhattan (o cityblock) e di Mahalanobis. Quest'ultima è una distanza non isotropa che prende in considerazione la correlazione tra le variabili. La distanza euclidea risulta essere la distanza ottimale per i due metodi migliori. Il confronto effettuato ha quindi permesso di individuare due dendrogrammi con la massima cophenetic correlation. In Fig. 2.5 è rappresentato il dendrogramma per il metodo 'average' e distanza euclidea con un esempio di taglio per 5 clusters.

Nelle linee guida dice di mostrare più dendrogrammi, magari ne aggiungo un altro.

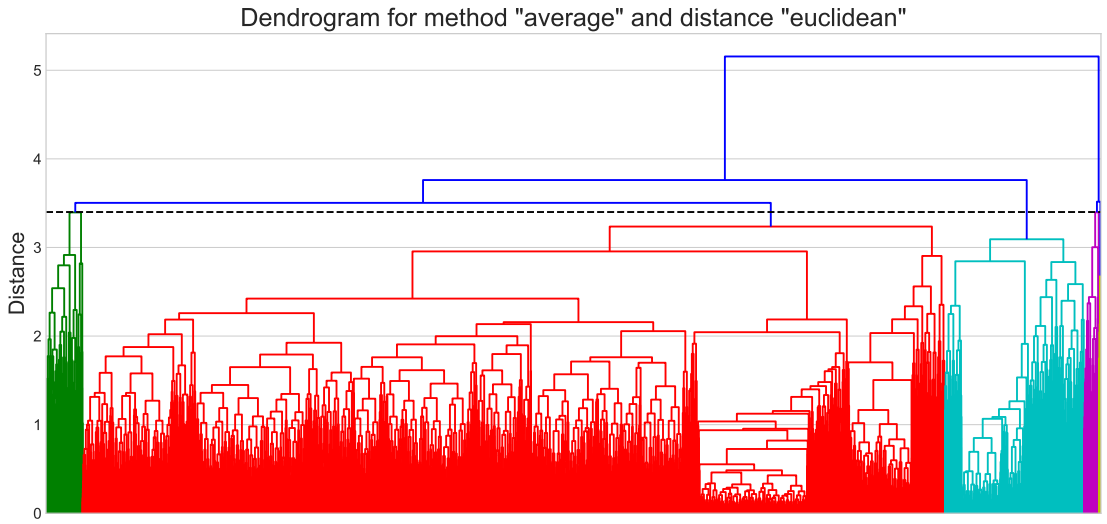


Figura 2.5: Dendrogramma per metodo 'average' e distanza euclidea, con esempio di taglio con 5 clusters, rappresentati dai differenti colori.

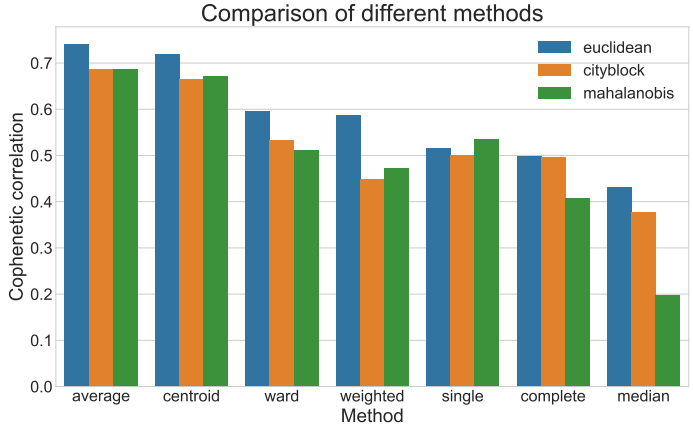


Figura 2.6: Confronto tra diversi metodi

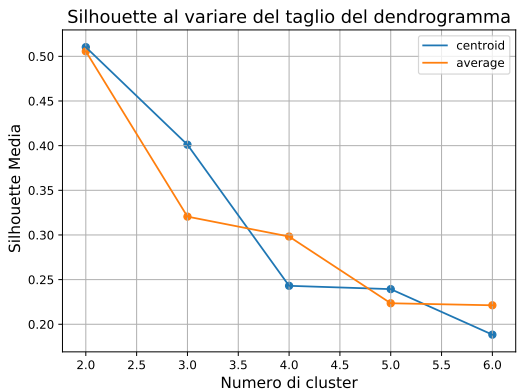


Figura 2.7: Silhouette media per i metodi 'average' e 'centroid' (con distanza euclidea) in funzione del numero di cluster

	cluster0	cluster1
average	14757	242
centroid	14775	224

	cluster0	cluster1	cluster2
average	12780	1977	242
centroid	13797	978	224

	cluster0	cluster1	cluster2	cluster3
average	30	212	1977	12780
centroid	224	978	3	13794

	cluster0	cluster1	cluster2	cluster3	cluster4
average	30	212	1977	12261	519
centroid	4	220	978	3	13794

Tabella 2.3: Numerosità nei cluster per rispettivamente 2,3,4,5 cluster, per i due metodi 'average' e 'centroid'.

Per confrontare i dendrogrammi tra di loro e con i risultati degli altri metodi di clustering è stata calcolata la silhouette media, tagliando i dendrogrammi ad un altezza tale da ottenere un numero di cluster compreso tra 2 e 6. I risultati sono mostrati in Fig 2.7 e mostrano che i due metodi anche se danno risultati simili possono presentare comunque scostamenti rilevanti, a parità di numero di cluster ottenuti con un taglio. Si può osservare ad esempio in Tab. 2.3 come la numerosità dei clusters vari in modo considerevole tra i due metodi.

In generale il clustering gerarchico risulta molto sensibile al metodo e alla metrica utilizzata.

2.4 Confronto tra metodi di clustering

TODO: confronto finale tramite silhouette dei vari clustering

2.5 Final evaluation of the best clustering approach and comparison of the cluster obtained

Dalle analisi condotte con le tre metodologie di clustering abbiamo rilevato che con la metodologia del DBSCAN non si riescono a ricavare delle informazioni rilevanti ai fini della ricerca in quanto presenta un cluster in cui vi sono la quasi totalità dei dipendenti che vogliamo andare ad analizzare. Il K-means invece offre l'opportunità di determinare tipi di dipendenti, in cui è possibile inoltre fare delle supposizioni, che potranno essere studiate nelle sezioni successive. Per lo hierarchical... . Quindi per questo dataset si è ritenuto più efficiente fare utilizzo della metodologia del kmeans.

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.

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

3.4 Discussion of the most interesting rules

Abbiamo inoltre rilevato delle rules che nonostante non siano frequenti sono comunque interessanti per l'analisi che stiamo portando avanti. Queste sono: - - Da queste possiamo ricavare che...

3.5 Use the most meaningful rules to replace missing values and evaluate the accuracy

Nel nostro dataset non sono presenti missing values quindi non è stata necessaria la valutazione delle più significative rules e la valutazione dell'accuratezza per rimpiazzare questi.

3.6 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

Al fine di costruire un *decision tree* con i record a nostra disposizione, abbiamo eseguito alcune manipolazioni prima della costruzione dell'albero vero e proprio: sia la variabile *Salary* che la variabile *Department* sono state convertite in variabili numeriche, per facilitare l'analisi dei record, inoltre, per fornire un *target* durante la costruzione dell'albero, la variabile *Left* è stata separata dal dataset fornito. Tutte le altre variabili sono state lasciate inalterate.

4.1 Learning of different decision trees

Durante la fase di costruzione vera e propia, abbiamo deciso di implementare due alberi distinti, differenziandoli in base alla metrica con cui sarebbe stato scelto lo *split* migliore. Abbiamo perciò costruito un albero basato sull'indice *Gini* come misura di impurità di un nodo, e un albero basato invece sull'uso dell'*Entropia*. Utilizzando questi due indici di impurità, al momento di decidere quale split imporre, viene selezionato quello con il *gain* maggiore. Inizialmente abbiamo costruito entrambi gli alberi senza dare un limite sul numero di samples che un nodo avrebbe dovuto contenere per poter essere diviso, successivamente però, per evitare di incorrere nel noto fenomeno dell'*overfitting*, abbiamo voluto esplorare le possibilità date dal consentire uno split a patto che almeno il 10% o il 20% del numero totale di data objects sia presente nel nodo da dividere. In questa fase iniziale abbiamo utilizzato l'intero dataset come base per il training dell'albero, nelle sezioni successive modificheremo questa assunzione, dividento i data objects forniti tra test set e training set in percentuali ben definite.

4.2 Decision trees interpretation

Facciamo adesso alcune considerazioni sugli alberi ottenuti applicando i parametri descritti in Sezione 4.1. Sia per l'albero basato sull'indice Gini che per quello basato sull'Entropia, la radice è rappresentata dalla variabile *Satisfaction Level*, che divide i data objects tra coloro che hanno uno score ≤ 0.465 e coloro che hanno uno score > 0.465. Questa decisione è indicativa del fatto che l'utilizzo di tale variabile consente un primo split più redditizio rispetto alle altre variabili a disposizione. Come era lecito aspettarsi, gli alberi ottenuti, non imponendo un vincolo sul numero di samples in ogni nodo prima dello split, hanno un'altezza considerevole, tale altezza decresce in modo significativo applicando i vincoli esposti in Sezione 4.1, arrivando ad ottenere un albero di altezza 13 per l'indice Gini e uno di altezza 11 per l'Entropia. In Figura 4.1 possiamo osservare i dettagli riguardanti le radici degli alberi di altezza 11, a destra, e 13, a sinistra. Al fine di ottenere questi alberi è stato limitato il numero di records minimo contenuto all'interno di un nodo al 20% del totale a disposizone nel dataset.

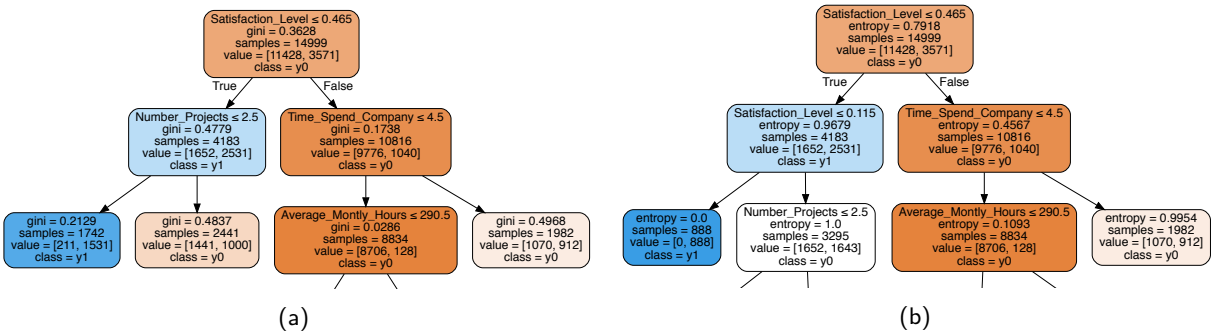


Figura 4.1: Dettaglio iniziale dell'albero di altezza 13 utilizzante Gini, in Figura 4.1a , e quello di altezza 11 utilizzante l'Entropia, in Figura 4.1b. La classe $y0$ corrisponde a una risposta negativa alla domanda: "L'impiegato lascerà l'azienda?", vale il contrario per la classe $y1$. Essendo l'intero albero di dimensioni proibitive, abbiamo deciso di rappresentare questo dettaglio.

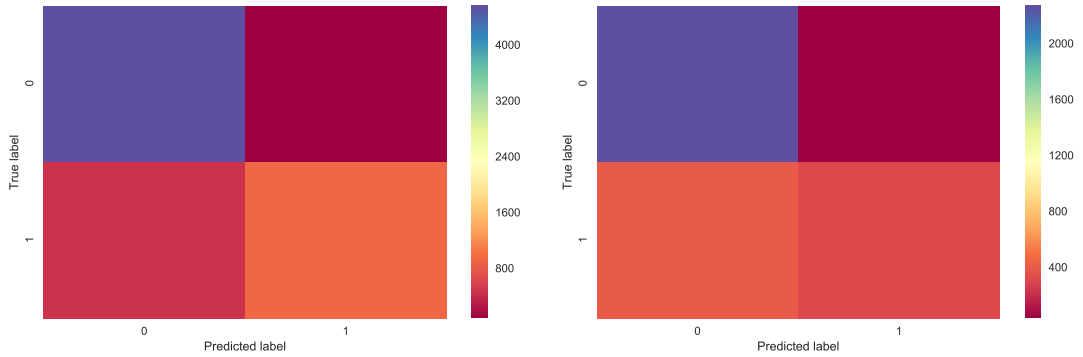
Come possiamo osservare, la colorazione dei vari nodi fornisce un indice visivo del grado di purezza dello split ottenuto, cominciando da uno split iniziale su tutto il dataset con un grado di purezza non molto alto (nel caso dell'indice Gini), si raggiunge successivamene negli split secondari, su porzioni ridotte del dataset, gradi di purezza più consistenti.

4.3 Decision trees validation with test and training set

Prendiamo quanto trattato nelle sezioni precedenti e passiamo adesso a descrivere le caratteristiche dei classificatori da noi costruiti. In Tabella 4.1 sono riportate le performance dei migliori 5 set di impostazioni per i classificatori. Come possiamo vedere dalla tabella, la nostra analisi è stata condotta prima prendendo l'intero dataset sia come training set che come test set, poi prendendone l'80% come training set e il restante 20% come test set e così via, modificando le percentuali fino ad

Criterion	Min Sample Split	Train Records	Test Records	Accuracy Training Set	Accuracy Test Set	Precision	Recall
Entropy	2	14999 (100%)	14999 (100%)	1.0	1.0	0.99	0.99
Gini	2	14999 (100%)	14999 (100%)	0.99	0.99	0.99	0.98
Entropy	2	8999 (60%)	6000 (40%)	0.99	0.98	0.96	0.94
Gini	2	11999 (80%)	3000 (20%)	0.99	0.98	0.95	0.95
Gini	2	10499 (70%)	4500 (30%)	0.99	0.98	0.95	0.95

Tabella 4.1: Tabella contentente le performance relative alle metriche di valutazione utilizzate per la classificazione tramite indice Gini ed Entropia. Min Sample Split rappresenta la minima percentuale di record rispetto al dataset totale che deve essere contenuta in ogni nodo (il valore 2 rappresenta il fatto che almeno 2 nodi devono essere presenti all'interno del nodo).



(a) (b)

Figura 4.2: In Figura 4.2a troviamo la confusion matrix relativa alla terza riga di Tabella 4.1, mentre in Figura 4.2b troviamo quella relativa alla quarta riga di Tabella 4.1.

utilizzare il 60% del dataset come training set e la restante parte come test set. La valutazione delle performance è stata condotta analizzando lo score ottenuto utilizzando le metriche *Accuracy*, *Precision* e *Recall*, apprese durante il corso.

4.4 Discussion of the best prediction model

Descriviamo adesso i risultati dei migliori classificatori ottenuti in quest'ultimo passo del report. Come era lecito aspettarsi, i primi due classificatori per score ottenuto nelle metriche utilizzate sono quelli costruiti utilizzando tutto il dataset come training set e successivamente come test set. Successivamente, il classificatore utilizzante l'Entropia come misura di impurità ha rivelato delle performance migliori rispetto a quello utilizzante Gini. In Figura 4.2a possiamo trovare la *confusion matrix* relativa a tale classificatore. Come possiamo vedere, la sezione relativa ai *true negatives* contiene il massimo dei records, come evidenziato dal colore, i *false negatives* e i *false positives* sono le due sezioni contenenti il numero minimo (tendente a 0) di records, facilmente prevedibile viste le alte performance di questo classificatore. La parte interessante della figura è rappresentata da quella dei *true positives*. Visto che il campione sulla quale è stato applicato questo classificatore è di 6000 records, e che tra questi data objects sono presenti per lo più impiegati che lavorano ancora all'interno dell'azienda, possiamo concludere che coloro che hanno lasciato l'azienda sono stati correttamente previsti dal classificatore. A scopo informativo, in Figura 4.2b abbiamo riportato la confusion matrix relativa alla quarta riga di Tabella 4.1, che, a differenza dell'altra matrice descritta, utilizza Gini come indice di impurità. In quest'ultimo classificatore, applicato su 3000 test records, possiamo osservare un incremento nella parte della matrice relativa ai false negatives, dovuta al calo delle performance complessive nelle metriche analizzate.