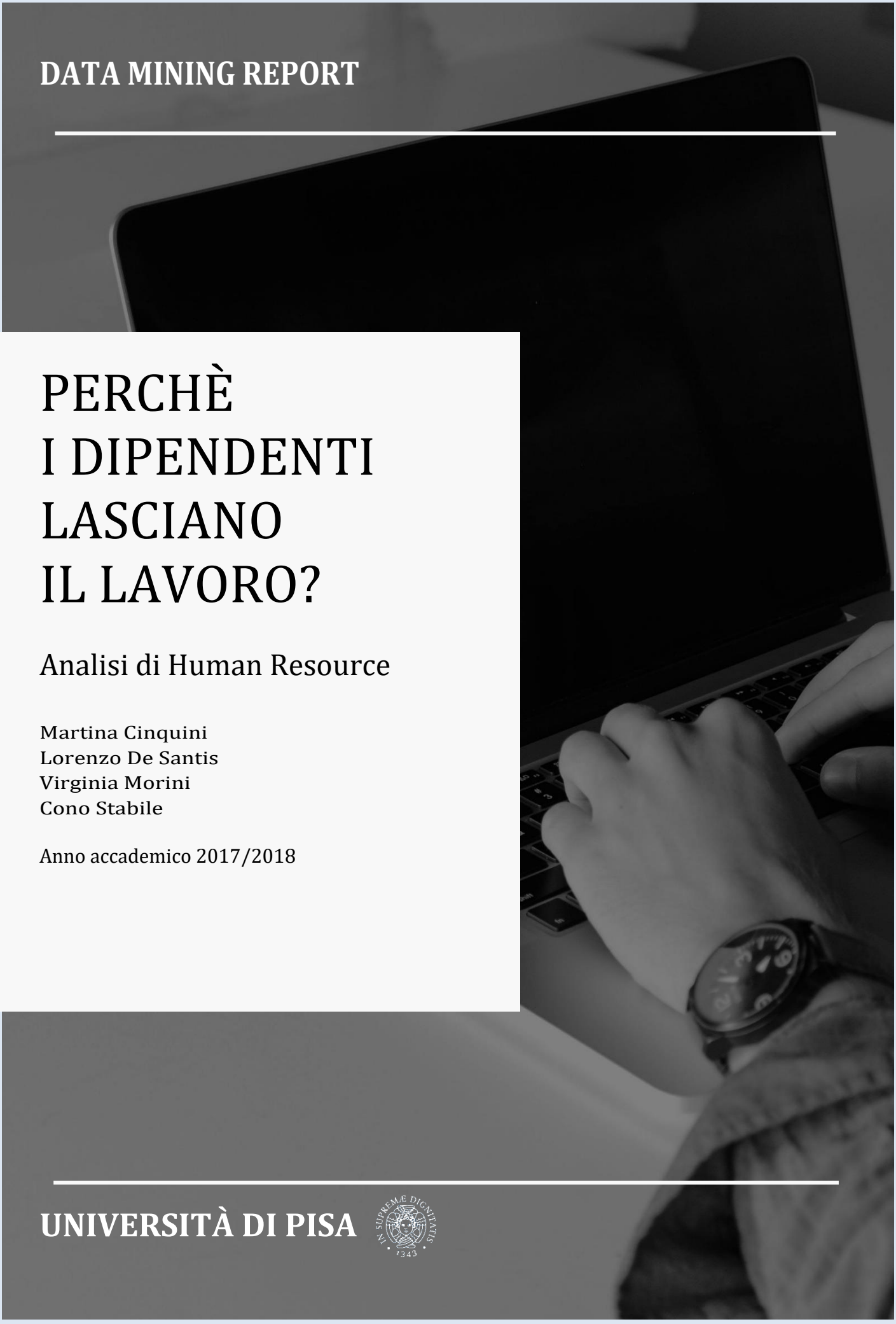


DATA MINING REPORT

PERCHÈ I DIPENDENTI LASCIANO IL LAVORO?

Analisi di Human Resource

Martina Cinquini
Lorenzo De Santis
Virginia Morini
Cono Stabile

Anno accademico 2017/2018

UNIVERSITÀ DI PISA

1. Data Understanding

Human Resources Analytics è il dataset utilizzato per la nostra analisi che si pone l'obiettivo di comprendere perché numerosi dipendenti dell'azienda lasciano il lavoro e, analizzando i loro comportamenti, predire chi potrebbe lasciarlo in futuro. Tale dataset contiene 14999 records descritti da 10 colonne di attributi.

1.1 Analisi degli attributi

Nella seguente tabella (Tabella 1) vengono descritti gli attributi raggruppandoli in base alla loro tipologia e indicando per ciascuno di essi il loro dominio.

Tabella 1. Descrizione e analisi degli attributi

Tipologia		Attributo	Descrizione	Dominio
Numerici	Continui	satisfaction_level	Indica il livello di soddisfazione di ogni dipendente	$[0.09,1] \cap \mathbb{R}$
		last_evaluation	Indica da quanto tempo è stata effettuata l'ultima valutazione delle prestazioni dei dipendenti in anni.	$[0.36,1] \cap \mathbb{R}$
	Discreti	number_project	Indica il numero di progetti completati a cui il dipendente ha collaborato.	$[2,7] \cap \mathbb{N}$
		average_montly_hours	Indica le ore mensili di lavoro di ogni dipendente	$[96,310] \cap \mathbb{N}$
		time_spend_company	Indica il numero di anni che il dipendente ha trascorso nell'azienda	$[2,10] \cap \mathbb{N}$
Categorici	Ordinali	salary	Indica la fascia di stipendio del dipendente	{low, medium, high}
	Nominali	sales	Indica il reparto in cui il dipendente lavora	{management, IT, product_mng, hr, sales, RandD, marketing, support, accounting, technical}
	Binari	Work_accident	Indica se il dipendente ha avuto un incidente sul posto di lavoro	{True, False}
		left	Indica se il dipendente ha lasciato il posto di lavoro o meno	{True, False}
		promotion_last_5years	Indica se il dipendente è stato promosso o meno negli ultimi cinque anni di lavoro	{True, False}

1.2 Analisi della qualità dei dati

La qualità del dataset è stata analizzata secondo i parametri di seguito riportati.

Accuratezza sintattica

Il dataset presenta alcuni errori sintattici così corretti:

- *number_project* è stato modificato in *number_projects* poiché i progetti svolti dai dipendenti possono essere più d'uno, risulta più corretto utilizzare la forma plurale;
- *average_montly_hours* è stato modificato in *average_monthly_hours* poiché 'monthly' era scritto in modo ortograficamente errato;
- *time_spend_company* è stato modificato in *time_spent_company* poiché tale attributo indica il tempo trascorso da ogni dipendente nella compagnia, risulta più corretto sostituire la forma presente del verbo 'spend' con quella passata 'spent';
- *Work_accident* è stato sostituito da *"work_accident"* per rispettare la coerenza del dataset in quanto tutti gli altri attributi sono scritti in minuscolo.

Abbiamo inoltre verificato che non fossero presenti errori di digitazione nei valori degli attributi di tipo stringa *sales* e *salary*.

Accuratezza semantica

Da un punto di vista semantico l'attributo *sales*, essendo un valore di tale attributo e non un descrittore di colonna, è stato modificato in *working_area,* in quanto i suoi valori indicano il reparto in cui lavorano i dipendenti. È possibile che il titolo incoerente sia legato ad un salvataggio non corretto in memoria del 15000° record il cui valore per l'attributo *working_area* potrebbe essere appunto 'sales'.

Completezza

I due attributi con valori continui *satisfaction_level* e *last_evaluation* sembrano normalizzati in quanto 1 è il valore massimo, ma il minimo non è 0 e anche nella descrizione dei metadati del database il range risulta [0,1]. Inoltre, per quanto riguarda l'attributo *last_evaluation*, i metadati specificano che la misura sia 'in anni', quindi ci sono due possibilità: i dati, in anni, sono stati normalizzati oppure indicano la frazione di anno trascorsa dall'ultima valutazione (in quest'ultimo caso, il valore limite 1 potrebbe significare che è obbligatoria almeno una valutazione all'anno).

Gestione degli Outliers e dei Missing Values

Al fine di individuare eventuali outliers, per ciascun attributo è stato generato il relativo Boxplot, che non ne ha evidenziato la presenza. Risulta però significativo il Boxplot relativo alla variabile *time_spent_company* (Figura 1) analizzato di seguito.

Il grafico mostra la presenza di outliers con valori compresi tra 6 e 10. Ma, considerando il loro significato, riteniamo sia possibile che all'interno di un'azienda ci sia un gruppo di membri anziani molto ristretto rispetto al totale dei dipendenti: pertanto non possiamo escluderli come valori anomali poiché potrebbero essere dati utili ai fini delle ulteriori analisi.

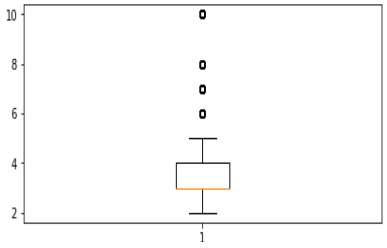


Figura 1. Box plot time_spent_company

Per quanto riguarda i missing values, all'interno del dataset non sono presenti valori nulli o assenti. È tuttavia possibile che, per l'attributo *satisfaction_level*, tali valori siano nascosti dal valore 0.1, data la sua alta e discostata frequenza rispetto ai valori simili e globali: una possibile ipotesi è che 0.1 fosse il valore di default nel questionario di valutazione.

1.3 Trasformazione e normalizzazione delle variabili

Al fine di eseguire correttamente il calcolo delle correlazioni e dei clustering, sono state apportate le seguenti modifiche alle variabili:

- Si è effettuata la trasformazione da stringhe a valori numerici del dominio dell'attributo *salary* che è stato mappato in {0, 0.5, 1};
- È stata effettuata la normalizzazione min-max degli attributi *satisfaction_level*, *last_evaluation*, *number_projects*, *average_monthly_hours* e *time_spent_company*. Il dominio delle seguenti variabili è stato quindi trasformato in un range [0,1].

1.4 Distribuzione delle variabili e analisi statistiche

Distribuzione attributi numerici

Si sono utilizzati istogrammi con curva gaussiana per mostrare la distribuzione dei valori di ogni attributo numerico del Dataset. Per scegliere il numero di bins ottimale per le variabili *satisfaction_level* (Figura 2a), *last_evaluation* (Figura 2b) e *average_monthly_hours* (Figura 3b) si è applicata la regola di Sturges; per *number_projects* (Figura 3a) e *time_spent_company* (Figura 3c) invece, avendo un range di valori ristretto, si è preferito effettuare un conteggio di tali valori, ottenendo rispettivamente 6 e 9 bins.

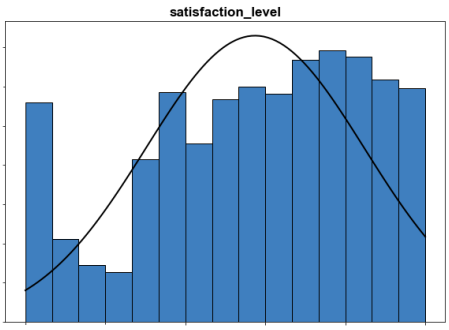


Figura 2a. Distribuzione satisfaction_level

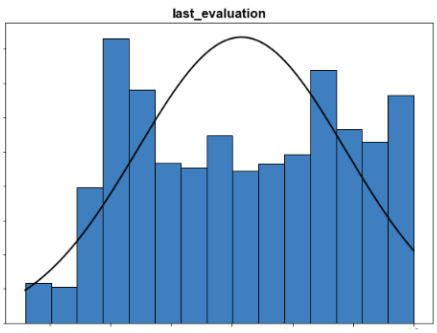


Figura 2b. Distribuzione last_evaluation

Tramite l'istogramma dell'attributo *satisfaction_level* (Figura 2a) è possibile osservare un picco di frequenza intorno a 0.1, valore che coincide con la moda, la quale si discosta di molto dalla media e dalla concentrazione principale dei valori (per i motivi già evidenziati nel paragrafo 'Gestione degli Outliers e dei Missing Values'). Inoltre, attraverso la gaussiana (Figura 2a) è possibile osservare quanto la distribuzione della variabile *satisfaction_level* si avvicini alla distribuzione normale. Nella Tabella 2 sono riportate le misure statistiche calcolate per le variabili numeriche continue:

Tabella 2. Misure statistiche delle variabili numeriche continue

Attributi	Media	Moda	Mediana	Deviazione Standard
satisfaction_level	0.61	0.1	0.64	0.28
last_evaluation	0.72	0.55	0.72	0.17

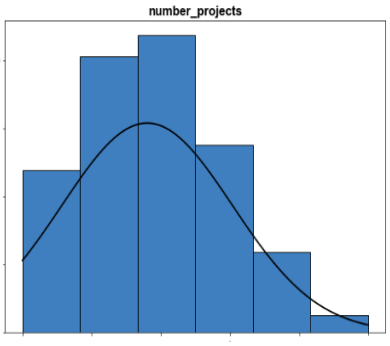


Figura 3a. Distribuzione number_projects

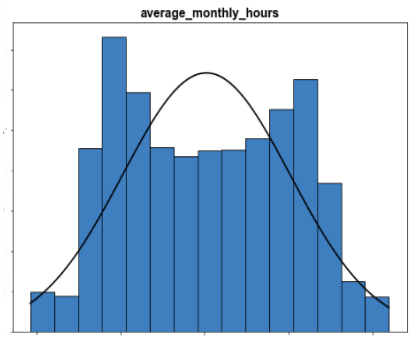


Figura 3b. Distribuzione average_monthly_hours

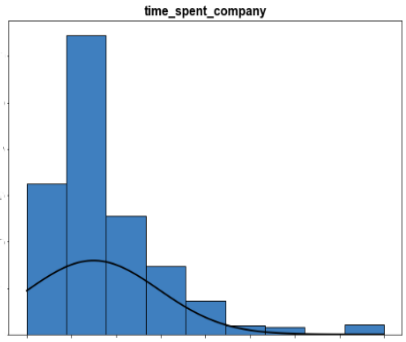


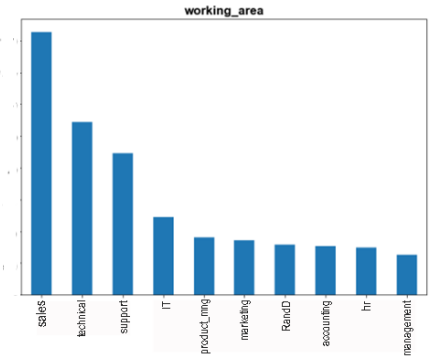
Figura 3c. Distribuzione time_spent_company

Le variabili discrete seguono perlopiù un andamento normale: in particolare la variabile average_monthly_hours sembra avere un comportamento bimodale (Figura 3b) mentre nella Figura 3c è possibile notare gli outliers che sfuggono dal profilo della gaussiana. Nella Tabella 3 sono riportate le misure statistiche calcolate per le variabili numeriche discrete.

Tabella 3: Misure statistiche delle variabili numeriche discrete

Attributi	Media	Moda	Mediana	Deviazione Standard
number_projects	3.80	4	4.0	1.23
average_monthly_hours	201.05	(135,136)	200.0	49.94
time_spent_company	0.72	3	3.0	1.46

Distribuzione attributi categorici



Dall'istogramma dell'attributo nominale *working_area* (Figura 4) si evince che il maggior numero di impiegati dell'azienda lavorano nel settore vendite, seguito dal settore tecnico e di supporto. Un numero minore di impiegati lavorano invece nel settore management, seguito dal settore risorse umane.

Figura 4: Distribuzione working_area

L'istogramma dell'attributo ordinale *salary* (Figura 5) presenta una distribuzione logaritmica dei dati. Tale risultato è piuttosto prevedibile se si pensa che all'interno di un'azienda lo stipendio rispecchia le responsabilità ripartite seguendo i criteri della piramide di Anthony[1].

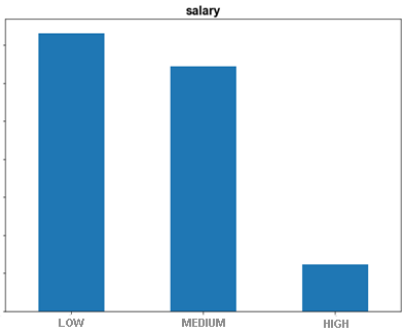


Figura 5. Distribuzione salary

Per gli attributi binari *left* (Figura 6a), *promotion_last_5years* (Figura 6b) e *work_accident* (Figura 6c) si sono realizzati grafici a torta avendo come istanze valori booleani. Risulta interessante evidenziare che solo il 2% degli impiegati ha ricevuto una promozione negli ultimi cinque anni e che quindi il dato potrebbe essere interessante solo in caso di particolari comportamenti per quella piccola frazione di dipendenti.

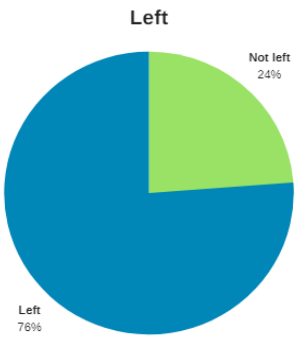


Figura 6a.
Distribuzione Left

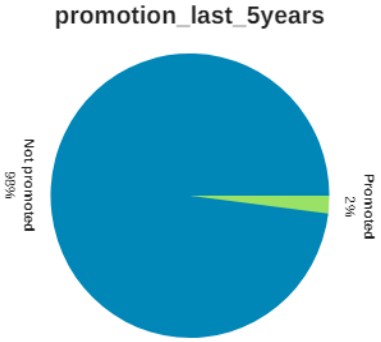


Figura 6b.
Distribuzione promotion_last_5years

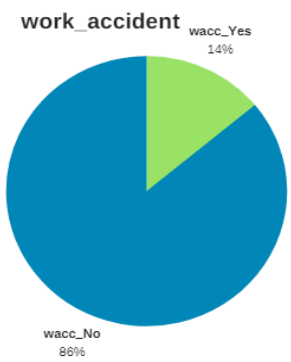


Figura 6c.
Distribuzione work_accident

1.5 Correlazioni tra variabili

Il calcolo delle correlazioni è stato effettuato escludendo gli attributi binari *left*, *promotion_last_5years*, *work_accident* e l'attributo nominale *working_area.* Si è scelto di utilizzare il coefficiente di correlazione di Pearson, poiché comparandone i risultati con quelli relativi al coefficiente di Spearman non si sono rilevate variazioni significative. Graficamente, tali risultati sono rappresentati tramite la mappa di calore in Figura 7.

Da tale mappa emergono valori di correlazione significativi che vengono di seguito analizzati in ordine decrescente:

[1] La piramide di Anthony è un sistema di classificazione gerarchico per l'organizzazione delle imprese.

1. *number_projects* e *average_monthly_hours* presentano il più alto valore di correlazione positivo: 0.417. Logicamente, all'aumentare del numero dei progetti a cui il dipendente collabora aumentano anche le sue ore lavorative.
2. *time_spent_company* e *number_projects* presentano un valore di correlazione positiva di 0.196. Logicamente, all'aumentare degli anni di lavoro all'interno dell'azienda aumenta il numero dei progetti a cui il dipendente ha collaborato.
3. *number_projects* e *satisfaction_level* presentano un valore di correlazione negativa di -0.142. Risulta interessante che all'aumentare del numero dei progetti a cui il dipendente collabora, diminuisce il suo livello di soddisfazione.
4. *time_spent_company* e *average_monthly_hours* presentano un valore di correlazione positiva di 0.127. Si evince che all'aumentare degli anni in cui il dipendente lavora nell'azienda aumentino anche le sue ore lavorative mensili.
5. *satisfaction_level* e *time_spent_company* presentano un valore di correlazione negativa di -0.100. Anche se tale valore è piuttosto basso, risulta significativo che all'aumentare degli anni passati dal dipendente all'interno dell'azienda diminuisca il suo livello di soddisfazione.

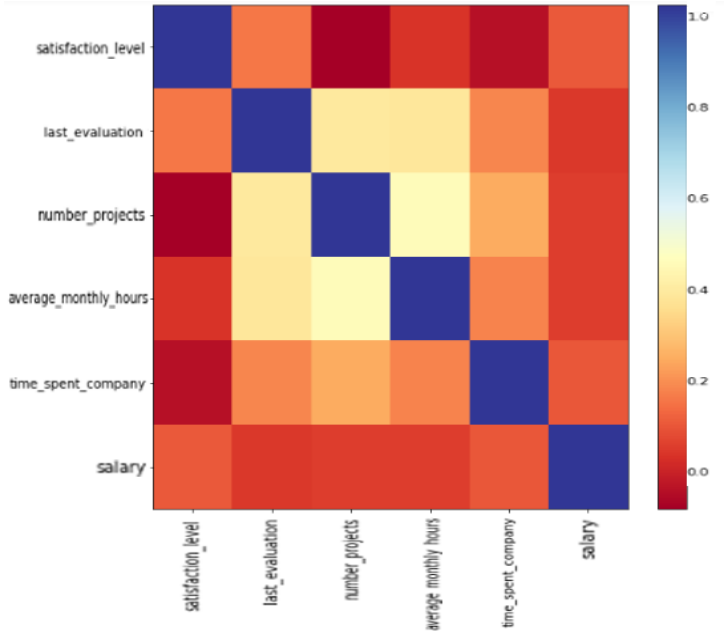


Figura 7. Mappa di calore delle correlazioni tra attributi

Come si evince dalla matrice di correlazione in cui non vi sono valori particolarmente alti, all'interno del dataset non sono presenti attributi ridondanti; pertanto si è deciso di non eliminare nessuna variabile.

2. Clustering

In questo capitolo si descrive l'applicazione delle tecniche di clustering sul dataset ottenuto nella fase di Data understanding escludendo gli attributi di tipo booleano (*work_accident, left, promotion_last_5years*) e categorico (*working_area, salary*) in quanto sul dominio normalizzato le distanze eccessive tra 0 e 1 avrebbero sparso i punti. Di conseguenza, tali attributi avrebbero avuto un carico eccessivo ai fini delle analisi. Quindi la feature selection considera i seguenti attributi: *satisfaction_level, last_evaluation, number_projects, average_monthly_hours, time_spent_company.*

2.1 K-means

Per stimare il parametro *k* con cui eseguire l'algoritmo di K-means si sono calcolate la SSE e la silhouette al variare di k su un range da 2 a 100.

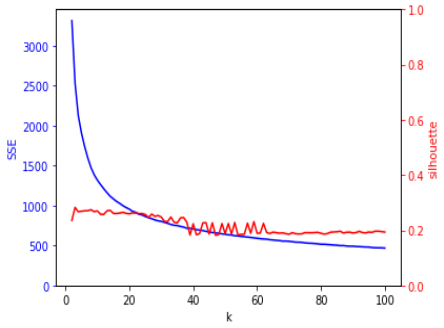


Figura 8: SSE e silhouette

Secondo quanto illustrato in Figura 8, sull'asse delle ascisse è presente il valore k e sull'asse delle ordinate il corrispondente valore di SSE e silhouette. Il grafico mostra l'andamento della curva ottenuta plottando i risultati dell'esecuzione dell'algoritmo. Notiamo che il cambio di pendenza è presente in corrispondenza di *k=10*. A partire dalla stima data dalla SSE, si è eseguito il K-means con valori $3 \leq k \leq 12$ utilizzando come metrica di distanza quella euclidea. Si è reiterato l'algoritmo 10 volte con nuove estrazioni di centroidi ad ogni step per evitare problemi legati alla scelta casuale dei centroidi iniziali. Infine, si sono utilizzati degli istogrammi per analizzare le distribuzioni dei singoli cluster rispetto al totale.

In relazione ai dati ottenuti, i clusters per *k=8* risultano essere i più significativi. Di seguito vengono analizzati, assegnando a ciascuno di essi un'ipotetica etichetta in base alle peculiarità che li contraddistinguono.

Clusters dei dipendenti che **hanno lasciato l'azienda**:

1. *Tirocinanti*: gruppo costituito da 2469 dipendenti che, svolgendo un numero esiguo di progetti e lavorando poche ore al mese, si ritengono piuttosto insoddisfatti e che hanno avuto una valutazione recente dall'azienda.
2. *Stakanovisti:* gruppo costituito da 1130 dipendenti che, portando a termine numerosi progetti e lavorando molte ore al mese, risultano decisamente insoddisfatti. Inoltre, hanno ricevuto una valutazione remota.
3. *Richiesti*: gruppo costituito da 2455 dipendenti che, svolgendo un numero medio di progetti e lavorando mensilmente un numero elevato di ore, risulta piuttosto soddisfatto. Hanno ricevuto una valutazione remota. Probabilmente, nonostante la soddisfazione, hanno lasciato l'azienda per un'offerta avuta da un'altra organizzazione.

Clusters dei dipendenti che **sono rimasti** nell'azienda:

1. *Fondatori*: gruppo costituito dai 601 dipendenti che hanno trascorso il maggior numero di anni all'interno dell'azienda. Gli stessi hanno realizzato un numero medio di progetti e hanno un numero elevato di ore mensili lavorative. Hanno un livello di soddisfazione medio-alto e hanno ricevuto una valutazione a distanza di tempo intermedio. Inoltre, molti sono stati promossi e hanno avuto incidenti.
2. *Veterani*: gruppo costituito da 1026 dipendenti che hanno trascorso un numero alto di anni all'interno dell'azienda. Essi hanno realizzato un numero significativo di progetti ma hanno un basso livello di soddisfazione.
3. *Reparto esecutivo neo-censito*: gruppo costituito da 2517 dipendenti che hanno ricevuto una valutazione recente e che lavorano da pochi anni nell'azienda. Nonostante lavorino mensilmente un numero elevato di ore e abbiano realizzato pochi progetti, risultano molto soddisfatti.
4. *Part-time*: gruppo costituito da 2223 dipendenti che lavorano da pochi anni nell'azienda. Hanno realizzato pochi progetti ma, a differenza del gruppo precedentemente descritto, hanno un numero di ore mensili relativamente basso. Inoltre, hanno ricevuto una valutazione recente.

5. *Reparto esecutivo*: gruppo costituito da 2578 dipendenti che hanno ricevuto una valutazione remota e lavorano da pochi anni nell'azienda. Pur lavorando a tempo pieno hanno realizzato pochi progetti ma risultano molto soddisfatti.

Si sono inoltre plottati i dati relativi a chi ha lasciato la compagnia in un grafico 3D (vedi Figura 9) prendendo come dimensioni *last_evaluation*, *satisfaction_level* e *average_monthly_hours*. Il risultato è molto significativo infatti si osservano i tre cluster sopra descritti (*Tirocinanti*, *Stakanovisti*, *Richiesti*) nettamente distinti e anche il valore della silhouette piuttosto alto indica l'alta similarità degli elementi appartenenti ad ogni cluster.

Infine, si è reiterato l'algoritmo escludendo dal dataset i 3 cluster relativi a chi ha lasciato l'azienda. Si è settato il parametro *k* del K-means a 5 e i risultati ottenuti risultano essere molto simili ai cluster individuati con *k=8* (esclusi ovviamente i 3 cluster di dipendenti che abbandonano).

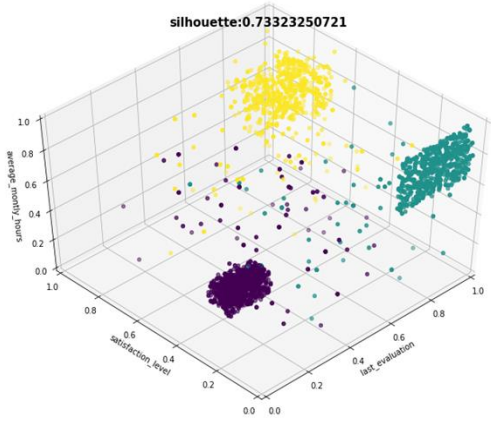


Figura 9: Cluster individuati con le caratteristiche dei dipendenti che hanno lasciato la compagnia

Di seguito, nella Tabella 4 sono riportate la media (ovvero le coordinate dei centroidi) e la deviazione standard relative ad ogni cluster individuato.

Tabella 4: Media e Deviazione Standard dei cluster ottenuti con K-means

Cluster	N° di elementi	sat_lev	last_ev	n_proj	avg_mnt_h	time_comp	w_acc	left	prom _5y
Tirocinanti	2455	0.42 ±0.08	0.53 ± 0.07	2.25 ±0.78	148.40 ±22.83	3.05 ±0.71	10%	60%	2%
Stakanovisti	1130	0.12 ±0.07	0.86 ±0.08	6.02 ±0.71	272.12 ±23.50	4.16 ±0.70	8%	80%	0%
Richiesti	2452	0.78 ±0.12	0.90 ±0.07	4.35 ±0.78	243.98 ±21.85	3.91 ±1.28	13%	36%	2%
Fondatori	600	0.68 ±0.18	0.71 ±0.15	3.62 ±0.90	199.78 ±44.34	8.14 ±1.51	21%	0%	9%
Veterani	1026	0.26 ±0.11	0.68 ±0.16	4.70 ±1.00	185.64 ±43.78	4.31 ±1.31	17%	0%	1%
Reparto Esecutivo Neo-censito	2517	0.71 ±0.13	0.61 ±0.09	3.60 ±0.90	242.37 ±22.42	2.88 ±0.85	17%	0%	2%
Part-time	2223	0.77 ±0.14	0.60 ±0.09	4.05 ±0.84	125.18 ±25.24	2.30 ±0.90	18.36 %	0.9 %	3%
Reparto Esecutivo	2596	0.72 ±0.15	0.87 ±0.08	3.44 ±0.79	176.40 ±31	2.86 ± 0.80	16%	1%	1%

2.2 DBSCAN

L'algoritmo è stato eseguito sul dataset utilizzato per il K-means, privo di variabili booleane e categoriche.
Per stimare i parametri ottimali dell'algoritmo si sono utilizzati i grafici del k-dist[2] usando come funzione di distanza quella euclidea. Si è preso in considerazione il range dei valori corrispondenti al gomito della curva che per epsilon corrisponde a $0.18 \leq \varepsilon \leq 0.23$ mentre per i minPoints corrisponde a $420 \leq minPts \leq 490$.
I parametri scelti che ci hanno permesso di ottenere dei clusters con maggiore densità sono: ε=0.210 e *minPts* = 480.
DBSCAN ha individuato la conformazione di 3 clusters (blu, verde, giallo) con densità sufficientemente alta e una quantità significativa di rumore.
Nel grafico 3D, in Figura 10, è possibile vedere come sono distribuiti i cluster sulle tre variabili prese in analisi:

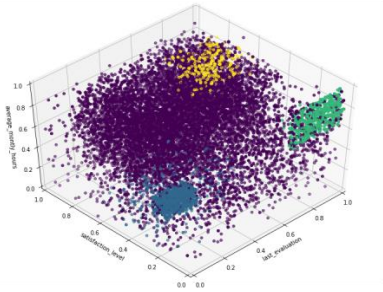


Figura 10: Cluster individuati con il DBSCAN

1. *Blu* è costituito dagli individui che hanno ricevuto una valutazione recente, che hanno un livello di soddisfazione medio-basso e che hanno svolto un numero basso di ore lavorative mensili;
2. *Verde* è costituito dagli individui che hanno ricevuto una valutazione remota, che hanno svolto un numero elevato di ore lavorative mensili e che hanno un basso livello di soddisfazione;
3. *Giallo* ha le stesse caratteristiche del cluster turchese ma differisce nel livello di soddisfazione che in questo caso è alto.

Data la notevole quantità di rumore si è voluto reiterare l'algoritmo solamente su quest'ultimo al fine di individuare nuovi clusters a densità minore rispetto ai primi tre. Si sono scelti, attraverso i grafici del k-dist, nuovamente i parametri di cui necessita l'algoritmo: ε = 0.185 e minPts = 200.
DBSCAN ha individuato altri due clusters costituiti dai dipendenti che non hanno lasciato l'azienda che presentano le seguenti caratteristiche:

1. *Rosso:* gruppo costituito dai dipendenti che lavorano da pochi anni nell'azienda con un basso numero di ore mensili. Hanno realizzato un numero esiguo di progetti, mostrano un livello di soddisfazione medio-basso e hanno ricevuto una valutazione remota;
2. *Grigio*: gruppo costituito dai dipendenti che hanno trascorso pochi anni nell'azienda lavorando mensilmente molte ore. Hanno realizzato un numero medio di progetti, mostrano un livello di soddisfazione medio e hanno ricevuto una valutazione remota.

Di seguito, nella Tabella 5, vengono riportati i valori della media e della deviazione standard calcolati su ogni cluster individuato.

Tabella 5: Media e Deviazione standard dei cluster individuati con DBSCAN

Cluster	N° elementi	sat_lev	last_ev	n_proj	avg_mnt_h	time_comp	w_acc	left	prom_ 5years
Blu	1896	0.42 ±0.06	0.51 ± 0.05	2.08 ±0.26	145.28 ±15.00	2.97 ±0.31	7.38%	80%	1%
Verde	800	0.10 ±0.01	0.87 ±0.05	6.18 ±0.45	276.29 ±19.64	4.08 ±0.45	4.87%	97%	0.3%
Giallo	671	0.81 ±0.06	0.92 ±0.06	4.85 ±0.35	244.90 ±16.12	5.11 ±0.46	4.76%	92%	0.2%
Rosso	2035	0.73 ±0.13	0.71 ±0.13	4±0	203.42 ±38.31	2.84 ±0.63	16.81 %	0%	3%
Grigio	1906	0.72 ±0.12	0.72 ±0.13	3±0	195.87 ±35.13	2.82 ±0.61	16.84 %	0%	1%
Noise	7691								

[2] P. Tan, M. Steinbach, V. Kumar, *Introduction to Data Mining*, Pearson Addison Wesley, 2006, p. 530

Ai fini di un'analisi più completa si è deciso di eseguire l'algoritmo DBSCAN anche sul dataset privo delle informazioni relative agli individui che hanno lasciato l'organizzazione (*left* a 1). Sono state effettuate diverse prove con vari settaggi dei parametri epsilon e minPoints.
Con $0.15 \leq eps \leq 0.19$ e $95 \leq minpts \leq 160$ si sono individuati tre cluster poco densi e con un'alta presenza di rumore. Si è deciso di abbassare minPoints fino a 35 al fine di ottenere maggiore densità e minor rumore, ma abbiamo visto che l'algoritmo divide solo sulla dimensione number_projects. Con $eps \geq 0.2$ e con $minpts \geq 195$ DBSCAN ha creato un unico cluster, mentre con $eps \leq 0.14$ e $minpts \geq 100$ l'algoritmo non ha creato nessun cluster, ma solo rumore. In conclusione, non sono state individuate informazioni rilevanti.

2.3 Hierarchical clustering

L'algoritmo è stato eseguito sul dataset utilizzato per il K-means e il DBSCAN, privo di variabili booleane e categoriche.
Si è scelto di utilizzare la formula euclidea per la distanza dato che altri metodi testati (manhattan, supremum) hanno condotto a risultati simili e pertanto non verranno riportati. Inoltre, come metodo di merge, si sono utilizzati due tipi di approccio di linkage ovvero singolo e completo.

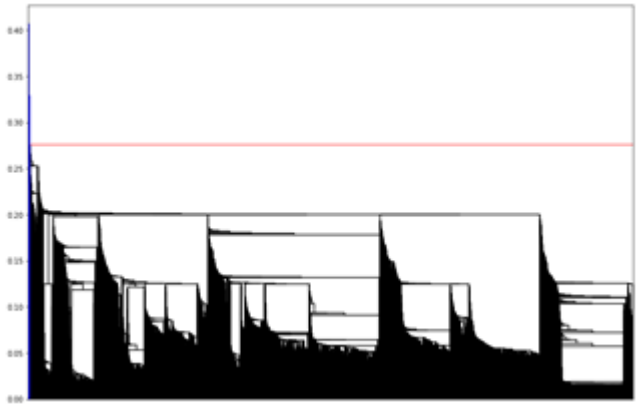


Figura 11: Single Linkage

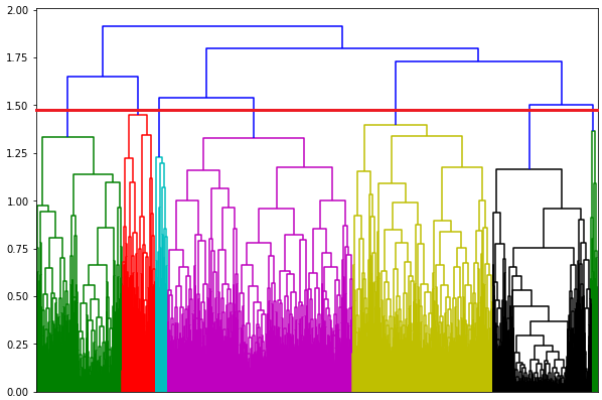


Figura 12: Complete Linkage

Il risultato ottenuto dal Single Linkage (Figura 11) è poco significativo, in quanto una soglia di ~ 0.30 sulla distanza genera un cluster con la quasi totalità dei dati e alcuni clusters di dimensioni non comparabili. Un eventuale taglio alternativo del dendrogramma porterebbe in ogni caso a disparità di diversi ordini di grandezza della dimensione dei clusters ottenuti. Attraverso l'analisi del Complete Linkage (Figura 12), invece, con una soglia ~ 1.48 si ottengono 8 clusters di dimensioni comparabili. Si è deciso di denominarli utilizzando un ordinamento numerico per non confonderli con le etichette assegnate ai clusters risultanti dagli algoritmi precedenti.
Clusters dei dipendenti che **sono rimasti nell'azienda**:

Cluster 1: gruppo costituito da 418 dipendenti insoddisfatti che hanno trascorso in media 4 anni nell'azienda, lavorando a tempo pieno e realizzando molti progetti.

Cluster 2: gruppo costituito da 477 dipendenti abbastanza soddisfatti che hanno trascorso in media 4 anni nell'azienda lavorando un numero significativo di ore mensili. Hanno realizzato molti progetti e hanno ricevuto una valutazione recente.

Cluster 4: gruppo costituito da 314 dipendenti che ha trascorso il maggior numero di anni nell'azienda lavorando mensilmente molte ore e realizzando pochi progetti. Hanno un livello di soddisfazione medio-alto e hanno ricevuto una valutazione intermedia.

Cluster 6: gruppo costituito da 3747 dipendenti soddisfatti che hanno trascorso pochi anni nell'azienda lavorando a tempo pieno e realizzando pochi progetti. Hanno ricevuto una valutazione recente.

Cluster 8: gruppo costituito da 180 dipendenti insoddisfatti che hanno trascorso molti anni nell'azienda lavorando mensilmente un numero elevato di ore e realizzando pochi progetti. Hanno un livello di soddisfazione basso, hanno realizzato un numero basso di progetti e hanno una distribuzione media delle ore mensili lavorative. Hanno ricevuto una valutazione recente.

Clusters dei dipendenti che **hanno lasciato l'azienda**:

Cluster 3: gruppo costituito da 2316 dipendenti insoddisfatti che hanno trascorso in media 3 anni e mezzo nella compagnia realizzando diversi progetti e lavorando mensilmente un eccessivo numero di ore. Inoltre, hanno ricevuto una valutazione remota.

Cluster 5: gruppo costituito dagli 4942 dipendenti soddisfatti che hanno trascorso in media 3 anni e mezzo nell'azienda realizzando un numero medio di progetti e lavorando mensilmente un numero elevato di ore. Hanno ricevuto una valutazione remota.

Cluster 7: gruppo costituito da 2605 dipendenti abbastanza soddisfatti che hanno trascorso nell'azienda in media meno di 3 anni realizzando pochi progetti e lavorando a tempo pieno. Inoltre, hanno ricevuto una valutazione recente.

La descrizione dei clusters ottenuti è riportata sotto forma tabellare (Tabella 6):

Tabella 6: Peculiarità clusters ottenuti con hierarchical

Cluster	N°elem	sat_lev	last_ev	n_proj	avg_mnt_h	time_comp	w_acc	left	Prom_5y
Cluster 1	418	0.29 ±0.12	0.67 ± 0.18	5.09 ±0.93	152.46 ± 29.53	4.32 ±1.33	16%	9%	2%
Cluster 2	477	0.46 ±0.17	0.56 ± 0.09	5.02 ± 0.71	242.09 ± 33.56	4.02 ±1.83	15%	9%	1%
Cluster 3	2316	0.31 ±0.22	0.85 ± 0.09	4.85 ± 1.40	250.88 ±34.36	3.70 ±1.08	12%	41%	1%
Cluster 4	314	0.69 ±0.18	0.74 ± 0.15	3.62 ± 0.84	199.35 ±41.31	8.88 ±1.41	18%	0%	8%
Cluster 5	4942	0.79 ±0.13	0.85 ±0.10	3.95 ±0.87	209.40 ±42.49	3.43 ±1.28	14%	19%	2%
Cluster 6	3747	0.75 ±0.15	0.59 ± 0.09	3.69 ±0.90	198.01 ±43.34	3.13 ±1.13	17%	1%	2%
Cluster 7	2605	0.44 ±0.10	0.55 ± 0.10	2.40 ±0.67	146.28 ±18.12	2.94 ±0.53	10%	59%	1%
Cluster 8	180	0.35 ±0.18	0.51 ± 0.09	2.79 ±0.65	192.94 ±61.04	5.88 ±2.08	22%	7%	6%

2.4 Confronto tra i clustering ottenuti

In conclusione, per ogni algoritmo di clustering si è ottenuto il seguente numero di clusters:

- K-means: 8
- DBSCAN: 5
- Hierarchical: 8

Nella Tabella 7 viene riportato il confronto tra i risultati ottenuti da K-means, DBSCAN e Hierarchical.

Tabella 7: Confronto tra i clustering ottenuti

K-means	DBSCAN	Hierarchical	Composizione cluster
Tirocinanti	*Blu*	*Cluster 7*	Ore lavorative mensili basse, livello di soddisfazione basso, valutazione recente.
Stakanovisti	*Verde*	*Cluster 3*	Ore lavorative mensili alte, livello di soddisfazione basso, valutazione remota.
Richiesti	*Giallo*	*Cluster 5*	Ore lavorative mensili alte, livello di soddisfazione alto, valutazione remota.
Reparto esecutivo	*Grigio*	*Cluster 6*	Ore lavorative mensili medie, livello di soddisfazione alto, valutazione intermedia, lavorano da pochi anni nell'azienda.
Part-time	*Cluster non trovato*	*Cluster non trovato*	Ore lavorative mensili basse, numero di progetti basso, lavorano da pochi anni nell'azienda, valutazione recente.
Veterani	*Cluster non trovato*	*Cluster 8*	Ore lavorative mensili alte, livello di soddisfazione basso, numero alto di progetti.
Fondatori	*Cluster non trovato*	*Cluster 4*	Ore lavorative mensili alte, livello di soddisfazione medio-alto, numero medio di progetti, lavorano da più tempo nell'azienda.
Reparto esecutivo neo-censiti	*Rosso*	*Cluster non trovato*	Ore lavorative mensili medie, livello di soddisfazione alto, numero medio-alto di progetti, lavorano da poco tempo nell'azienda.
Cluster non trovato	*Cluster non trovato*	*Cluster 1*	Ore lavorative mensili medio-basso, livello di soddisfazione basso, numero alto di progetti, lavorano da circa 4 anni e mezzo nell'azienda.
Cluster non trovato	*Cluster non trovato*	*Cluster 2*	Ore lavorative mensili alte, livello di soddisfazione medio, numero alto di progetti, lavorano da abbastanza tempo nell'azienda.

È interessante notare che il cluster *Fondatori*, ottenuto con K-means, non è ritrovabile con DBSCAN poiché, essendo costituito da outliers, gli elementi del cluster sono isolati e di conseguenza non sono sufficientemente densi.

2.5 Valutazione del miglior clustering

Basandoci sia sulla silhouette (Tabella 8) sia sulla semantica dei clusters ottenuti, abbiamo valutato che il miglior clustering per questo dataset sia dato dall'algoritmo K-Means con $k=8$, poiché i clusters ottenuti sono ben separati ed è stato possibile interpretarli ed assegnar loro un senso logico, consistente con ciò che accade in una vera organizzazione.

Tabella 8: Silhouette dei clusters

Algoritmi di clustering	Silhouette
K-means	0.274
DBSCAN	0.074
Hierarchical	0.021

3. Association Rules Mining

3.1 Estrazione e descrizione dei pattern più significativi

L'estrazione dei pattern è avvenuta mediante l'algoritmo "A Priori". Prima di eseguire l'algoritmo si è deciso di scartare gli attributi *promotion_last_5years* e *work_accident* per quei record in cui il loro valore era 0, in quanto non sarebbero stati informativi e sarebbero risultati matematicamente inutili nella generazione degli itemset, data l'alta frequenza del valore 0.
Si è utilizzata la discretizzazione secondo la legge di Sturges (15 bins per 14999 records) per gli attributi *satisfaction_level*, *last_evaluation* e *average_monthly_hours*. Trovando solo pattern con basso supporto, e quindi poco indicativi, si è deciso di accoppiare anche a due a due i valori di *time_spent_company* e *number_projects* e di abbassare il numero di bins da 15 a 10.
Sono stati così estratti 67 patterns con supporto superiore al 10% del totale (~1500 conteggi), tutti completi e di cui 33 massimali. In particolare, 23 di questi contengono informazioni significative correlate al licenziamento del dipendente. Di seguito, nella Tabella 9, vengono descritti dettagliatamente tali pattern.

Tabella 9. Descrizione pattern frequenti

Itemset	Supporto	Descrizione

`('138.80-160.20_A', '0_L')`
`('160.20-181.60_A', '0_L')`
`('181.60-203.00_A', '0_L')`
`('203.00-224.40_A', '0_L')`
`('224.40-245.80_A', '0_L')`
`('245.80-267.20_A', '0_L')`
	11.5%
	11.1%
	10.6%
	10.5%
	10%
	10.6%
		Un totale del 64 % dei dipendenti che ha un numero di ore mensili lavorative compreso tra 138.80-267.20 e non lascia l'organizzazione.

`('0.54-0.64_S', '0_L')`
`('0.64-0.73_S', '0_L')`
`('0.73-0.82_S', '0_L')`
`('0.82-0.91_S', '0_L')`
`('0.91-1.00_S', '0_L')`
	11.5%
	11.4%
	11.4%
	10.4%
	10.7%
		Un totale del 55% dei dipendenti che ha un livello di soddisfazione alto, quindi compreso tra 0.54-1 e non lascia l'organizzazione.

`('support', '0_L')`	11.2 %	Percentuale dei dipendenti che lavora nel settore supporto e non lascia l'azienda.
`('technical', '0_L')`	13.5 %	Percentuale dei dipendenti che lavora nel settore tecnico e non lascia l'azienda.
`('1_L', 'low')`	14.5 %	Percentuale di dipendenti che ha lasciato l'organizzazione e che ha un salario esiguo
`('1_L', '4-5_T')`	11.5%	Percentuale di dipendenti che ha lasciato l'organizzazione e che ha trascorso un numero di anni compreso tra 4 e 5 in questa.
`('1_L', '2-3_N', '2-3_T')`	10.4%	Percentuale di dipendenti che ha lasciato l'organizzazione, con un numero di progetti compreso tra 2 e 3 e ha trascorso un numero medio di anni compreso tra 2 e 3 in questa.
`('4-5_T', '0_L')`	15.4%	Percentuale dei dipendenti che ha trascorso un numero di anni all'interno dell'organizzazione compreso tra 4 e 5 e che non l'ha lasciata.
`('sales', '4-5_N', '0_L')`	11%	Percentuale dei dipendenti che ha operato nel reparto vendite e ha realizzato un numero di progetti compreso tra 4 e 5. Non ha lasciato l'organizzazione
`('sales', '2-3_T', '0_L')`	14.4%	Percentuale dei dipendenti che ha operato nel reparto vendite, ha trascorso un numero di anni all'interno dell'organizzazione compreso tra 2 e 3. Non ha lasciato l'organizzazione
`('2-3_N', 'medium', '2-3_T', '0_L')`	10.8%	Percentuale dei dipendenti che ha svolto un numero di progetti compreso tra 2 e 3 e che percepisce uno stipendio medio. Ha trascorso un numero di anni compreso tra 2 e 3 all'interno dell'azienda e non l'ha lasciata.

('2-3_N', 'low', '2-3_T', '0_L')	10.6%	Percentuale dei dipendenti che ha svolto un numero di progetti compreso tra 2 e 3 e che percepisce uno stipendio basso. Ha trascorso un numero di anni compreso tra 2 e 3 all'interno dell'azienda e non l'ha lasciata.
('medium', '4-5_N', '2-3_T', '0_L')	12.4%	Percentuale dei dipendenti che ha svolto un numero di progetti compreso tra 4 e 5 e che percepisce uno stipendio medio. Ha trascorso un numero di anni compreso tra 2 e 3 all'interno dell'organizzazione e non l'ha lasciata.
('4-5_N', 'low', '2-3_T', '0_L')	13.5%	Percentuale dei dipendenti che ha svolto un numero di progetti compreso tra 4 e 5 e che percepisce uno stipendio basso. Ha trascorso un numero di anni compreso tra 2 e 3 all'interno dell'organizzazione e non l'ha lasciata.
('0.36-0.45_S','2-3_N', '2-3_T')	10.8%	Percentuale dei dipendenti che ha un livello di soddisfazione basso compreso tra 0.36 e 0.45, che ha svolto un numero di progetti compreso tra 2 e 3 e ha trascorso un numero di anni all'interno dell'azienda compreso tra 2 e 3.
'0.49-0.55_E','2-3_T')	12.8%	Percentuale dei dipendenti che ha ottenuto una valutazione intermedia compresa tra 0.49 e 0.55 e ha trascorso un numero di anni all'interno dell'azienda compreso tra 2 e 3.
('technical', '2-3_T')11.66	11.7%	Percentuale dei dipendenti che lavora nel settore tecnico e ha trascorso un numero di anni all'interno dell'organizzazione compreso tra 2 e 3.
('138.80-160.20_A','2-3_N', '2-3_T')	10.4%	Percentuale dei dipendenti che ha un numero di ore mensili lavorative compreso tra 138.80 e 160.20. Ha svolto un numero di progetti compreso tra 2 e 3 e ha trascorso un numero di anni all'interno dell'azienda compreso tra 2 e 3.
('4-5_T', 'medium')	11.3%	Percentuale di dipendenti che ha trascorso un numero di anni compreso tra 4 e 5 all'interno dell'azienda e che percepisce uno stipendio medio.
('4-5_T', '4-5_N')	13.6%	Percentuale di dipendenti che ha trascorso un numero di anni compreso tra 4 e 5 all'interno dell'azienda e che ha svolto un numero di progetti compreso tra 4 e 5.
('4-5_T', 'low')	14%	Percentuale di dipendenti che ha trascorso un numero di anni compreso tra 4 e 5 all'interno dell'azienda e che percepisce uno stipendio basso.
('sales', '2-3_N')	12%	Percentuale dei dipendenti che lavora nel reparto vendite e che ha svolto un numero di progetti compreso tra 2 e 3.
('sales', 'medium')	11.8%	Percentuale dei dipendenti che lavora nel reparto vendite e che percepisce uno stipendio medio.
('sales', 'low')	14%	Percentuale dei dipendenti che lavora nel reparto vendite e che percepisce uno stipendio basso.

Inoltre, nel sottoinsieme composto da 3569 dipendenti che hanno lasciato l'azienda si sono estratti 19 itemsets (vedi Tabella 10), di cui 11 massimali, con un supporto relativo oltre il 20% (~700 conteggi) corrispondente a circa il 5% sul totale dei dati.

Tabella 10. Descrizione pattern frequenti dei dipendenti che hanno lasciato l'azienda

Itemset	Supporto	Descrizione
('126.00-144.40_A','2-3_T','2-3_N') ('144.40-162.80_A','2-3_T','2-3_N')	5.1% 4.9%	10% è la percentuale di chi ha lavorato nell'azienda per un periodo compreso tra 2 e 3 anni con un intervallo di ore mensili compreso tra 126.00 e 162.80 e che ha svolto un numero di progetti compreso tra 2 e 3.
('126.00-144.40_A', '2-3_N') ('144.40-162.80_A', '2-3_N')	5.1% 5.7%	10,8% è la percentuale di chi ha svolto un numero di ore lavorative mensili che ricoprono un intervallo tra 126.00 e 162.80 e che ha realizzato un numero di progetti compreso tra 2,3.
('126.00-144.40_A', '2-3_T') ('144.40-162.80_A', '2-3_T')	5,20% 5,01%	10,2% è la percentuale di chi ha svolto un numero di ore lavorative mensili che ricoprono un intervallo tra 126.00 e 162.80 e che hanno lavorato nell'azienda per un periodo compreso tra 2 e 3 anni.
('0.34-0.42_S', '2-3_T', '2-3_N')	6,4%	6,4% è la percentuale di chi ha lavorato nell'azienda per un periodo compreso tra 2 e 3, ha realizzato un numero di progetti compreso tra 2 e 3 e ha un livello di soddisfazione compreso tra 0.34 e 0.42.
('6-7_N', '0.09-0.17_S', '4-5_T')	5,6%	5,6% è la percentuale di chi ha lavorato nell'azienda per un periodo compreso tra 4 e 5 anni, ha realizzato un numero di progetti compreso tra 6 e 7 e ha un livello di soddisfazione compreso tra 0.09 e 0.17.
('4-5_N', '4-5_T')	5,1%	Percentuale di chi ha lavorato nell'azienda per un periodo compreso tra 4 e 5 anni e che ha svolto 4-5 progetti.
('4-5_T','low')	6.8%	Percentuale di chi ha lavorato nell'azienda per un periodo compreso tra 4 e 5 anni e ha un salario esiguo.
('2-3_T','2-3_N', 'low')	6,3%	6,3% è la percentuale di chi ha lavorato nell'azienda per un periodo compreso tra 2 e 3 e ha svolto un numero di progetti compreso tra 2 e 3 e ha un salario esiguo.

3.2 Estrazione e discussione delle association rules più significative

Data la necessità di costruire un classificatore a regole per l'attributo *left*, si sono estratte le regole più significative in termini di confidenza e lift contenenti informazioni su quell'attributo.
Sono state estratte 21 regole aventi l'attributo *left* con valore a 1 come conseguente tramite l'algoritmo "A Priori" con min_support = 2% (~300 conteggi) e confidence = 96%.
Tutte le regole, senza essere applicato nessun filtro, hanno un lift superiore a 4, ma di queste 21, solo 5 sono risultate non ridondanti e riportate in Tabella 11 al fine di fornire un'analisi dettagliata.

Tabella 11. Descrizione regole significative

Rules X→Y	Confidence	Lift	Descrizione
('288.60-310.00_A',) ---> 1_L	100%	4.20	Chi ha un numero di ore lavorative compreso tra 288-310 allora lascia l'azienda
('0.09-0.18_S', '6-7_N', '4-5_T', 'low') ---> 1_L	96.6%	4.06	Chi ha un livello di soddisfazione basso, ha realizzato 6/7 progetti, lavora all'interno dell'azienda da quattro o cinque anni e percepisce uno stipendio basso, allora lascia l'organizzazione.

('0.36-0.45_S', '0.49-0.55_E', '138.80-160.20_A', '2-3_N', 'low', '2-3_T') ---> 1_L	99.7%	4.19	Chi ha un livello di soddisfazione medio, e ha un numero di ore lavorative medio, e ha realizzato 2-3 progetti, e lavora all'interno dell'azienda da due-tre anni e percepisce uno stipendio basso, di conseguenza lascia l'organizzazione.
('0.36-0.45_S', '138.80-160.20_A', 'sales', '2-3_N') ---> 1_L	96.3%	4.05	Chi ha un livello di soddisfazione medio e che lavora nel settore vendite ricoprendo un numero di ore lavorative mensili compreso tra 138.80 e 160.20 e che ha realizzato due-tre progetti, allora lascia l'azienda.
('0.36-0.45_S', '138.80-160.20_A', 'sales', '2-3_T') ---> 1_L	96%	4.03	Chi ha un livello di soddisfazione medio e che lavora nel settore vendite ricoprendo un numero di ore lavorative mensili compreso tra 138.80 e 160.20 e che fa parte dell'azienda da due-tre anni, allora lascia quest'ultima.

Per l'attributo *left* con valore 0 sono state estratte 24 regole con min_support = 2% e confidence = 99.7% tra le quali si è scelto di descriverne 18 (vedi Tabella 12) selezionando solamente le regole non ridondanti.
Il lift ottenuto risulta essere piuttosto basso ossia ~ 1.3. Per questa ragione, si è provato anche ad abbassare la soglia di confidenza fino al 60%, ma senza buoni risultati, in quanto il lift delle regole ottenute non supera ~ 1.5.

Tabella 12. Analisi regole riguardanti i dipendenti che sono rimasti nell'azienda

Rules (X → Y)	Confidence	Lift	Descrizione
('0.36-0.42_E') ---> 0_L	100%	1.31	Chi ha ricevuto una valutazione intermedia allora non lascia l'azienda.
('96.00-117.40_A',) ---> 0_L	100%	1.31	Chi ha un numero di ore mensili lavorative tra 96 e 117.40, allora non lascia l'azienda.
('8-9_T',) ---> 0_L	100%	1.31	Chi fa parte della compagnia otto-nove anni, allora non lascia l'organizzazione. È da evidenziare una certa affinità con la suddetta regola e il cluster dei fondatori individuato con il K-means.
('high', '4-5_N', '2-3_T') ---> 0_L	100%	1.31	Chi ha realizzato 4-5 progetti e ha percepito un salario alto, e fa parte dell'azienda da 2-3 anni, di conseguenza non ha lasciato quest'ultima.
('0.62-0.68_E', 'medium', '2-3_T') ---> 0_L	99.7%	1.31	Chi ha ricevuto una valutazione intermedia, percependo un salario medio e facente parte dell'azienda da 2-3 anni, allora non ha lasciato l'organizzazione.
('0.68-0.74_E', '2-3_N', '2-3_T') ---> 0_L	99.8%	1.31	Chi ha ricevuto una valutazione quasi remota, ha realizzato 2-3 progetti, facente parte dell'azienda da 2-3 anni, allora non ha lasciato l'organizzazione.
('181.60-203.00_A', 'medium', '4-5_N') ---> 0_L	100%	1.31	Chi ha un numero medio di ore lavorative e che ha realizzato 4-5 progetti percependo un salario medio, allora non lascia l'azienda.
('0.91-1.00_S', 'sales', '2-3_T') ---> 0_L	100%	1.31	Chi ha un livello di soddisfazione alto e opera nel settore vendita, e che fa parte della compagnia 2-3 anni, allora non lascia l'azienda.
('0.91-1.00_S', '2-3_N', 'medium') ---> 0_L	100%	1.31	Chi ha un livello di soddisfazione alto e che ha realizzato 2-3 progetti percependo un salario medio, allora non lascia l'azienda.
('0.91-1.00_S', '2-3_N', 'low') ---> 0_L	100%	1.31	Chi ha un livello di soddisfazione alto e che ha realizzato 2-3 progetti percependo un salario basso, allora non lascia l'azienda.

('0.91-1.00_S', '2-3_N', '2-3_T') ---> 0_L	100%	1.31	Chi ha un livello di soddisfazione alto e che ha realizzato 2-3 progetti e fa parte della compagnia da 2-3 anni, allora non lascia l'azienda.
('0.91-1.00_S', 'medium', '2-3_T') ---> 0_L	100%	1.31	Chi ha un livello di soddisfazione alto percependo uno stipendio medio e fa parte della compagnia da 2-3 anni, allora non lascia l'azienda.
('0.91-1.00_S', '4-5_N', 'low', '2-3_T') ---> 0_L	100%	1.31	Chi ha un livello di soddisfazione alto e ha realizzato 4-5 progetti percependo un salario basso, e che fa parte della compagnia da 2-3 anni, allora non lascia l'azienda.
('0.64-0.73_S', '4-5_N', 'low', '2-3_T') ---> 0_L	100%	1.31	Chi ha un livello di soddisfazione medio-alto e ha realizzato 4-5 progetti percependo un salario basso, e che fa parte della compagnia da 2-3 anni, allora non lascia l'azienda.
('0.82-0.91_S', '4-5_N', 'low', '2-3_T') ---> 0_L	99.7%	1.31	Chi ha un livello di soddisfazione alto e ha realizzato 4-5 progetti percependo un salario basso, e che fa parte della compagnia da 2-3 anni, allora non lascia l'azienda.
('0.73-0.82_S', 'medium', '4-5_N', '2-3_T') ---> 0_L	100%	1.31	Chi ha un livello di soddisfazione medio-alto e ha realizzato 4-5 progetti percependo un salario medio, e che fa parte della compagnia da 2-3 anni, allora non lascia l'azienda.
('support', 'medium', '4-5_N', '2-3_T') ---> 0_L	100%	1.31	Chi lavora nel settore supporto e ha realizzato 4-5 progetti percependo un salario medio e che fa parte della compagnia da 2-3 anni, allora non lascia l'azienda.
('0.49-0.55_E', 'medium', '4-5_N') ---> 0_L	100%	1.31	Chi ha ricevuto una valutazione intermedia e ha realizzato 4-5 progetti percependo un salario medio, allora non lascia l'azienda.

Inoltre, ai fini di un'analisi più completa, si sono esaminate anche le regole ottenute senza considerare l'attributo *left*, estraendo così 26 regole di cui 9 non ridondanti con min_support = 2% e confidence = 95%. Di seguito, vengono riportate le regole raggruppate per conseguente:

- `('0.42-0.49_E', '138.80-160.20_A', '2-3_T') ---> 2-3_N`. Questa regola appartiene a un gruppo costituito da altre 7 regole con conseguente 2-3_N aventi come antecedenti gli stessi attributi, ma con intervalli diversi. La confidence di ciascuna regola è ~ 98 %, mentre il lift è ~ 2.30. Quindi possiamo dedurre che chi ha un numero di ore lavorative medio, ha un salario basso e lavora all'interno dell'azienda da 2-3 anni, allora ha realizzato 2-3 progetti.
- `('288.60-310.00_A',) ---> 4-5_T`, ha come conseguente 4-5_T con confidence = 95.1% e lift = 3.54. Concludiamo che chi ha un numero di ore lavorative compreso tra 288.60 e 310 allora è dipendente dell'azienda da 4-5 anni.

3.3 Classificatore rule-based

Avendo solo due valori nella classe bersaglio, è possibile utilizzare solo uno dei due set di regole, in modo da riconoscere uno solo dei due valori. In particolare, abbiamo optato per il set di regole con *left=1* per il loro alto valore di lift e di confidence e per la frequenza più bassa dell'attributo all'interno del dataset. Abbiamo quindi costruito un classificatore che utilizza le 5 regole descritte in Tabella 11 per individuare i dipendenti che lasciano l'azienda e che classifica tutti quelli non corrispondenti alle regole come dipendenti fedeli. Il classificatore ottenuto ha riportato i valori delle metriche mostrati in Tabella 13.

Tabella 13. Metriche di valutazione della performance del classificatore

Precision	Accuracy	Recall	F-measure
97,5%	83,1%	30,1%	46,0%

La percentuale di accuracy risultante è dell' 83% e supera del 7% quella del classificatore banale che assegna sempre 0 ovvero la sua frequenza per l'attributo *left*. Inoltre, si può notare un ottimo livello di precisione da ricondursi agli alti valori di confidence e di lift delle regole.
Infine, per valutare il classificatore si è scomposto il dataset in Training Set (⅔) e Test Set (⅓). Per effettuare questa operazione, è stato sufficiente estrarre, ogni 3 record, un record da inserire nel Test Set, mentre gli altri due sono stati destinati al Training Set. Abbiamo dovuto utilizzare questo metodo poiché estrarre come Test Set gli ultimi 5000 o i primi 5000 record portava a risultati molto sbilanciati. A parità di confidence, sono state estratte dal Training Set 5 regole molto simili a quelle precedenti, riportate in Tabella 14. Non ci stupisce, pertanto, che anche le performance del nuovo classificatore, valutate sul Test Set e riportate in Tabella 15, siano simili.

Tabella 14. Regole ottenute dal training set

Rules (X → Y)	Confidence	Lift	Descrizione
`('288.60-310.00_A',) ---> 1_L`	100%	4.20	Chi svolge un elevato numero di ore mensili, allora lascia l'azienda
`('0.09-0.18_S', '6-7_N', '4-5_T', 'low') ---> 1_L`	96.1%	4.04	Chi ha un livello di soddisfazione basso, e che ha realizzato un numero di progetti tra 6, 7 percependo uno stipendio basso, allora lascia l'organizzazione.
`('0.36-0.45_S', '0.49-0.55_E', '138.80-160.20_A', '2-3_N', '2-3_T') ---> 1_L`	99.3%	4.17	Chi ha un livello di soddisfazione medio, e ha un numero di ore lavorative medio, e ha realizzato 2,3 progetti, e lavora all'interno dell'azienda da due-tre anni e ha ricevuto una valutazione intermedia di conseguenza lascia l'organizzazione.
`('0.36-0.45_S', '0.49-0.55_E', '2-3_N', 'low', '2-3_T') ---> 1_L`	97.8%	4.11	Chi ha un livello di soddisfazione medio, e ha realizzato 2,3 progetti, e lavora all'interno dell'azienda da due-tre anni e percepisce uno stipendio basso e che ha ricevuto una valutazione intermedia, di conseguenza lascia l'organizzazione.
`('0.36-0.45_S', '138.80-160.20_A', '2-3_N', 'medium', '2-3_T') ---> 1_L`	96.6%	4.06	Chi ha un livello di soddisfazione medio, e ha realizzato 2,3 progetti, e lavora all'interno dell'azienda da due-tre anni e percepisce uno stipendio medio svolgendo un numero di medio ore lavorative, di conseguenza lascia l'organizzazione.

Tabella 15. Metriche di valutazione della performance del classificatore rule-based

	Performance	
	Training Set	**Test Set**
Precision	97,4%	98,4%
Accuracy	84,7%	84,9%
Recall	36,9%	37,3%
F-Measure	53,5%	54,1%

Le regole ideali ottenute come descrizione dei 3 clusters individuati estraendo solo i record con *left* = 1 sarebbero: `('0.09-0.11_S', '0.77-0.97_E', '242-310_A', '6-7_N', '4-5_T') --> '1_L';('0.36-0.46_S', '0.45-0.57_E', '126-160_A', '2_N', '3_T') --> '1_L'; ('0.72-0.92_S', '0.81-1_E', '215-275_A', '4-5_N', '5-6_T') --> '1_L'` e sono molto simili a quelle trovate dall'algoritmo (porterebbero ad una accuracy del 97%).
Per quanto riguarda i missing values, non essendo presenti nel database, non è possibile usare le regole per stimarli. Se ci fossero stati, avremmo potuto stimarli creando un classificatore a regole per l'attributo del valore mancante, in modo analogo a come è stato creato il classificatore per l'attributo *left*.

4. Classificazione

Per costruire un classificatore sull'attributo *left*, si è innanzitutto escluso l'attributo *working_area* (oltre all'attributo *left*) poiché l'algoritmo utilizzato (libreria Scikit-learn) non è in grado di gestire attributi categorici e si è mappato l'attributo *salary* sul dominio {0,0.5,1}.
Per la valutazione delle performance, si sono poi considerati gli stessi training set e test set adottati per il classificatore a regole. Si è operata una grid search per il modello ad albero decisionale singolo basata su:

- criterion: [entropy, gini]
- max_depth: [2-20, None]
- min_impurity_decrease: [1e-6,5e-6,1e-7,5e-7,0]

Si è mantenuto sempre min_sample_split=2 e min_samples_leaf=1, perché si è visto che alzandoli si ottenevano performance strettamente peggiori.
I risultati migliori, ottenuti con una cross-validation con 10 folds, sono mostrati nelle Tabelle 16 e 17:

- Con parametri: max_depth:17, criterion:'entropy', min_impurity_decrease:1e-06

Tabella 16: Metriche di valutazione della performance del classificatore per il modello 1

	Performance	
	Training Set	**Test Set**
Precision	99,9%	94,9%
Accuracy	99,8%	98,2%
Recall	99,6%	99,6%
F-Measure	99,8%	95,6%

- Con parametri: max_depth:19, criterion:'entropy', min_impurity_decrease: 5e-07

Tabella 17: Metriche di valutazione della performance del classificatore per il modello 2

	Performance	
	Training Set	**Test Set**
Precision	99,8%	95,2%
Accuracy	99,8%	98,3%
Recall	99,0%	96,0%
F-Measure	99,4%	95,6%

Nonostante l'alta accuracy sul training set, che potrebbe far pensare all'overfitting, si hanno ottimi risultati anche sul test set. Questo ci porta a due conclusioni: la suddivisione training set/test set eseguita non è sbilanciata e i modelli ottenuti sono validi.

Si è inoltre operata una grid search per il modello a foresta di alberi decisionali basata su:

- criterion: [entropy, gini]
- max_depth: [2-20, None]
- min_impurity_decrease: [1e-6,5e-6,1e-7,5e-7,0]
- n_estimators: [10,15,20,25,30]

- min_samples_split: [2-51]
- min_samples_leaf: [2, 51]
- class_weight: [balanced, None, {0: 0.3, 1: 0.7}]

Si è mantenuto sempre bootstrap=True, perché si è visto che portandolo a False le performance peggioravano strettamente.
I risultati migliori sono mostrati nelle Tabelle 18 e 19:

- Con parametri max_depth:20, criterion:'entropy', min_impurity_decrease:1e-06, n_estimators:10, min_samples_split:8, min_samples_leaf:2, class_weight None, si è ottenuta una Mean validation del 98,2 %.

Tabella 18: Risultati ottenuti per la grid search per il modello a foresta di alberi decisionali per il modello 3

	Performance	
	Training Set	**Test Set**
Precision	99,4%	99,1%
Accuracy	98,6%	98,0%
Recall	95,0%	92,7%
F-Measure	97,2%	95,8%

- Con parametri max_depth:15, criterion:'entropy', min_impurity_decrease:1e-07, n_estimators:20, min_samples_split:8, min_samples_leaf:4, class_weight balanced, si è ottenuta una Mean validation del 98,1 %.

Tabella 19: Risultati ottenuti per la grid search per il modello a foresta di alberi decisionali per il modello 4

	Performance	
	Training Set	**Test Set**
Precision	99,0%	98,6%
Accuracy	98,9%	98,1%
Recall	96,5%	93,6%
F-Measure	97,7%	96,1%

Poiché i tre gruppi di dipendenti che hanno abbandonato il lavoro sono ben divisi, gli alberi singoli e gli alberi delle foreste ottenute sono tutti abbastanza simili fra loro, almeno per i primi 4-5 livelli: si differenziano invece per quanto riguarda la gestione dei pochi punti "rumorosi". Perciò le performance ottenute sono ottime e simili per tutti i modelli costruiti.

Il miglior modello per la predizione risulta quindi essere il secondo albero decisionale singolo (Figura 13) data l'alta accuracy, superiore sia alle foreste trovate, che al classificatore a regole. Crediamo che questo risultato sia dato dall'alta compattezza e definizione dei 3 cluster di dipendenti che abbandonano il posto di lavoro e dalla relativa assenza di rumore.

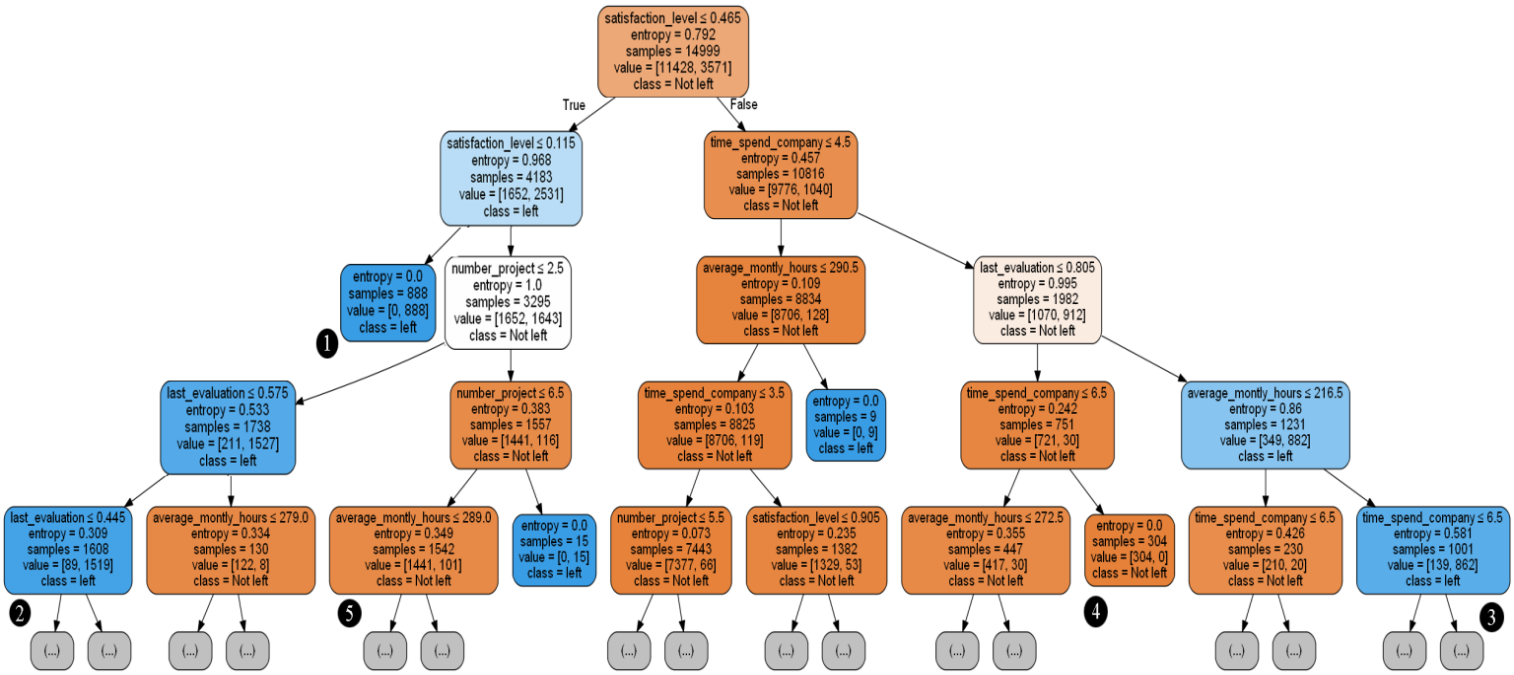


Figura 8: I primi 4 livelli dell'albero decisionale scelto

Si può notare come alcuni dei nodi dell'albero decisionale siano molto simili ad alcuni cluster, trovati in modo totalmente indipendente, risultanti dall'esecuzione degli algoritmi di clustering. Tale somiglianza rafforza la validità di entrambi i risultati, portandoci a pensare che, effettivamente, siano presenti nei dati i comportamenti descritti. I cluster individuati sono contrassegnati da un numero:

1. È il cluster degli *Stakanovisti*
2. È il cluster dei *Tirocinanti*
3. È il cluster dei *Richiesti*
4. È il cluster dei *Fondatori*
5. È il cluster dei *Veterani*

5. Conclusioni

Concludendo, si può notare che molti dipendenti lasciano l'azienda dopo circa 3-4 anni di lavoro: si può ipotizzare che tale comportamento sia dovuto all'impossibilità dei lavoratori di fare carriera all'interno dell'organizzazione, con l'eccezione di una cerchia molto ristretta di dipendenti. Pertanto, dopo un certo periodo, la maggior parte dei membri dell'azienda preferisce migrare verso altre realtà.