Human Resources Analytics
Data Mining, A.A. 2017/2018

Carlo Alessi
Francesco Cariaggi
Leonardo Cariaggi

5 gennaio 2018

Indice

1 Data understanding

In questa sezione si illustra il processo di *data understanding* attuato sul dataset. In particolare, nel paragrafo 1.1 si descrive la semantica e il tipo dei dati. Nel paragrafo 1.2 si discute la distribuzione degli attributi e si presentano alcune statistiche. Nel paragrafo 1.3 si valuta la qualità dei dati (rilevazione dei valori mancanti e degli *outlier*). Nel paragrafo 1.4 si descrivono quindi le trasformazioni applicate ai valori degli attributi. Infine, nel paragrafo 1.5, si ragiona sulla correlazione tra coppie di attributi e sull'eventuale eliminazione di attributi ridondanti (in base ai risultati ottenuti).

1.1 Semantica dei dati

Ogni riga del dataset contiene le informazioni di un singolo impiegato dell'azienda. La tabella 1.1 mostra la semantica e il tipo del valore di ogni colonna del dataset:

Nome dell'attributo	**Descrizione**	**Tipo**	Dominio
satisfaction_level	Livello di soddisfazione dell'impiegato	Numerico, continuo	$[0, 1] \subseteq \mathbb{R}$
last_evaluation	Ultima valutazione delle performance dell'impiegato	Numerico, continuo	$[0, 1] \subseteq \mathbb{R}$
number_project	Numero di progetti completati dall'impiegato nel periodo di lavoro	Numerico, discreto	$\mathbb{N}^+$
average_montly_hours	Numero medio di ore trascorse dall'impiegato ogni mese sul posto di lavoro	Numerico, discreto	$\mathbb{N}^+$
time_spend_company	Numero di anni trascorsi dall'impiegato nell'azienda	Numerico, discreto	$\mathbb{N}^+$
Work_accident	Indica se l'impiegato ha avuto un incidente sul posto di lavoro o meno	Numerico, categorico	$\{0, 1\}$
left	Indica se l'impiegato ha lasciato il posto di lavoro o meno	Numerico, categorico	$\{0, 1\}$
promotion_last_5years	Indica se l'impiegato ha ottenuto una promozione negli ultimi 5 anni o meno	Numerico, categorico	$\{0, 1\}$
sales	Dipartimento per il quale l'impiegato lavora	Stringa, non ordinale	$\{'sales', 'accounting', 'hr', 'technical', 'support', 'management', 'IT', 'product_mng', 'marketing', 'RandD'\}$
salary	Fascia di salario nella quale rientra l'impiegato	Stringa, ordinale	$\{'low', 'medium', 'high'\}$

Tabella 1.1: Semantica dei dati

1.2 Distribuzione degli attributi e statistiche

La tabella 1.2 mostra invece alcune statistiche riguardanti i valori degli attributi nel dataset: per ogni variabile sono riportate, in ordine, il numero di record per cui il valore non è mancante, la media, la deviazione standard, il minimo, i quartili (primo, secondo e terzo) e infine il massimo. Gli attributi booleani sono stati inclusi solo per avere un'idea della frequenza dei diversi valori in punti percentuali.
Di seguito si mostrano, separatamente, le distribuzioni degli attributi categorici e numerali del dataset. Precisamente, la figura 1.1 mostra la distribuzione degli attributi categorici: nell'asse delle ascisse sono riportati i valori assunti da ogni variabile, mentre sulle ordinate compare la frequenza dei singoli valori.

	satisfaction level	last evaluation	number project	average montly hours	time spend company	Work accident	left	promotion last 5years
count	14999	14999	14999	14999	14999	14999	14999	14999
mean	0.613	0.716	3.803	201.05	3.498	0.145	0.238	0.021
std	0.249	0.171	1.233	49.943	1.46	0.352	0.426	0.144
min	0.09	0.36	2	96	2	0	0	0
25%	0.44	0.56	3	156.0	3	0	0	0
50%	0.64	0.72	4	200.0	3	0	0	0
75%	0.82	0.87	5	245.0	4	0	0	0
max	1.0	1.0	7	310	10	1	1	1

Tabella 1.2: Statistiche del dataset

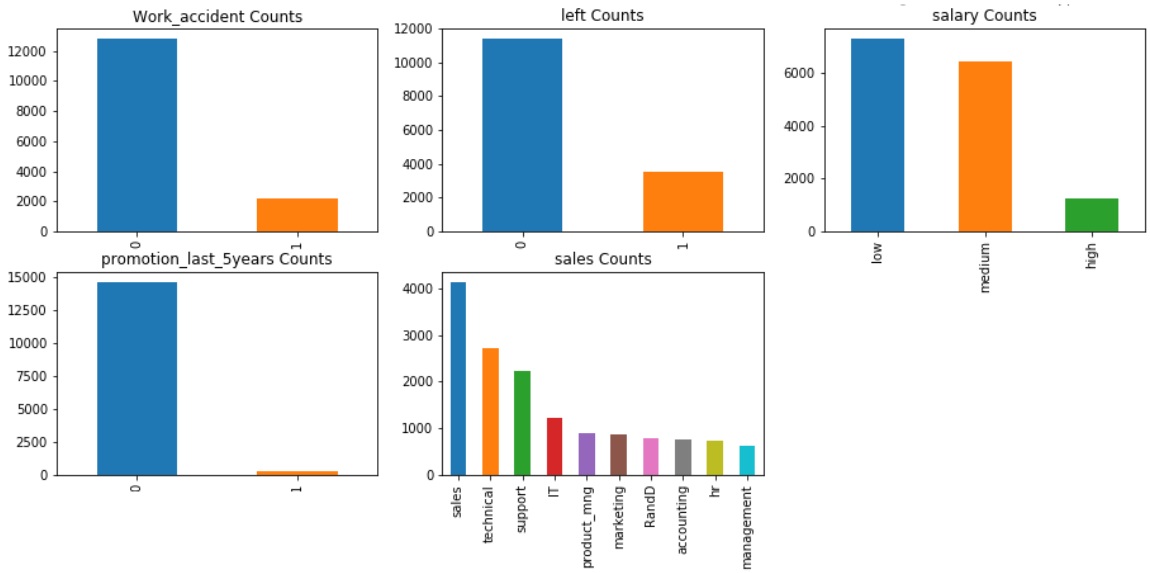


Figura 1.1: Distribuzione delle variabili categoriche

La figura 1.2, invece, riporta gli istogrammi che riassumono la distribuzione delle variabili numeriche del dataset. Le feature `number_project` e `time_spend_company`, in questa sezione, sono considerate numeriche (sebbene in altre fasi dell'analisi esse assumano il ruolo di variabili categoriche).

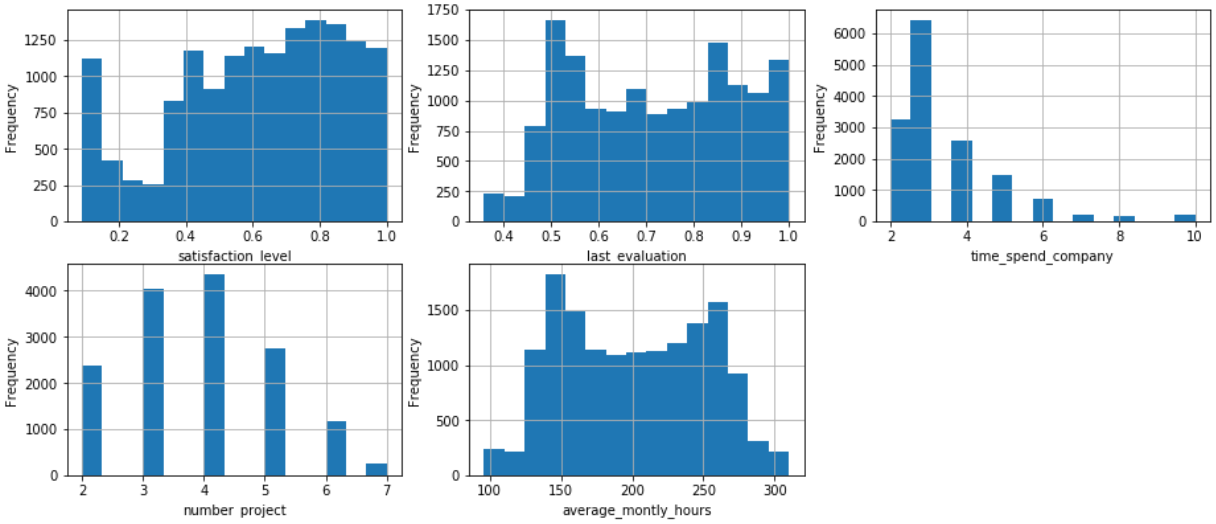


Figura 1.2: Distribuzione delle variabili numeriche

Osservazioni addizionali

La figura 1.3 mostra la distribuzione degli attributi numerici rispetto al valore assunto dalla variabile `left`:

- `last_evaluation`: tra gli impiegati che hanno lasciato l'azienda, la distribuzione assume due picchi tra 0,4 - 0,6 (intuitivamente, una valutazione bassa) e tra 0,8 - 1 (valutazione alta). Gli altri impiegati, invece, hanno valori distribuiti più o meno uniformemente.
- `satisfaction_level`: si vede chiaramente che gli impiegati che non hanno lasciato l'azienda sono quelli per i quali `satisfaction_level` assume per lo più valori alti. Gli impiegati che si sono licenziati, invece, mostrano dei picchi nei valori tra 0 e 0,4.
- `time_spend_company`: qui si nota che gli impiegati non lasciano quasi mai il lavoro nei primi due anni, bensì prendono la loro decisione tra il terzo e il quinto anno.
- `average_montly_hours`: i due picchi tra gli impiegati che hanno lasciato l'azienda dimostrano che essi abbandonano il lavoro o perché lavoravano troppo o perché lavoravano troppo poco (nel secondo caso, le ragioni principali sono da ricercare anche in altri fattori).
- `number_project`: qui vediamo che la maggior parte degli impiegati che lascia il posto di lavoro ha svolto soltanto due progetti (ciò significa che probabilmente hanno preso la loro decisione dopo aver incontrato le prime difficoltà).

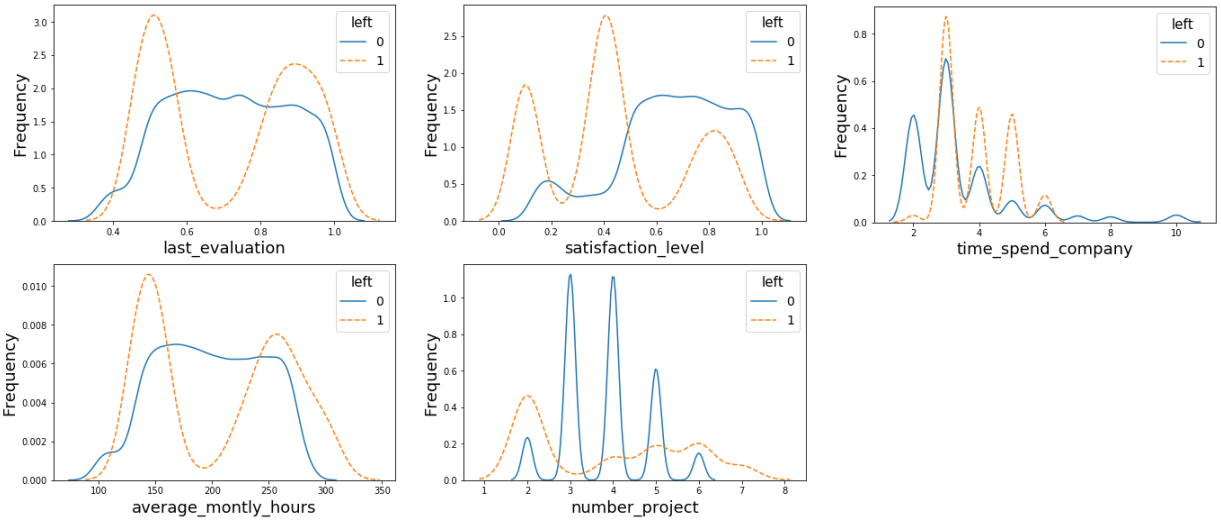


Figura 1.3: Distribuzione delle variabili numeriche rispetto al valore di left

La figura 1.4 mostra invece come il tempo complessivo trascorso in azienda dagli impiegati che lavorano nel dipartimento *'management'* è superiore alla media generale.

Infine, la figura 1.5 testimonia un comportamento ragionevole degli impiegati: un reddito elevato è spesso una buona ragione per rimanere. Tra coloro che hanno un salario alto, infatti, sono pochi quelli che alla fine decidono di abbandonare il proprio posto di lavoro. Tra gli altri impiegati, invece, la percentuale di abbandono è decisamente più alta.

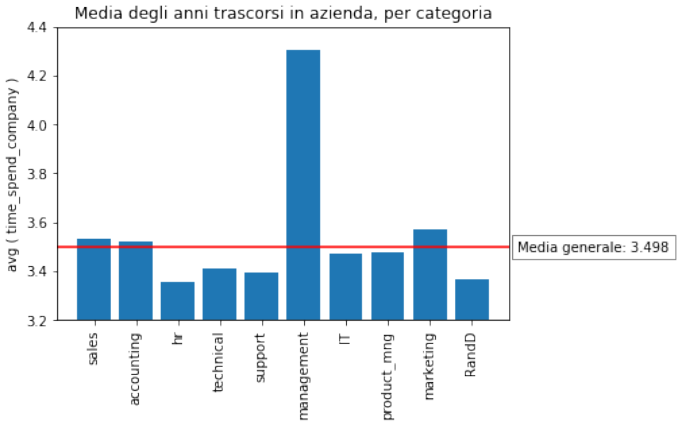


Figura 1.4: Tempo medio trascorso in azienda, in base al dipartimento

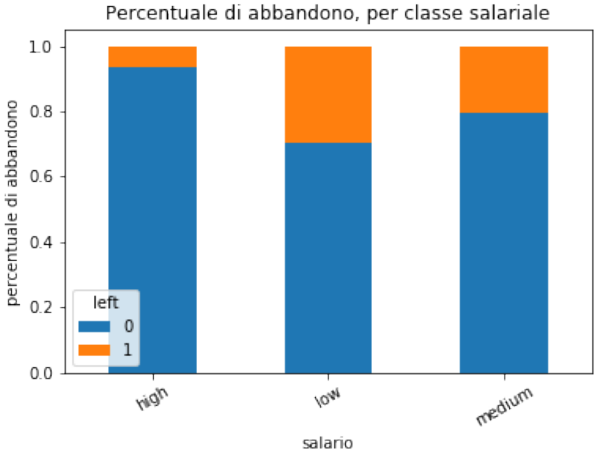


Figura 1.5: Percentuale di abbandono, per fascia di reddito

1.3 Valutazione della qualità dei dati

Come è possibile intuire dalla tabella 1.2, nel dataset non ci sono valori mancanti. Inoltre, i singoli valori di ogni feature sono coerenti con i domini specificati nella tabella 1.1: non sono dunque presenti errori dal punto di vista della sintassi dei dati (*Syntactic Accuracy*).

Trattandosi di un dataset simulato, assumiamo anche l'accuratezza semantica dei dati (i dati rispecchiano una situazione reale e sono *unbiased*), la completezza (coincidenza tra ciò che l'analisi richiede e ciò che i dati effettivamente raccontano) e la *Timeliness* (nessun ritardo nella disponibilità dei dati).

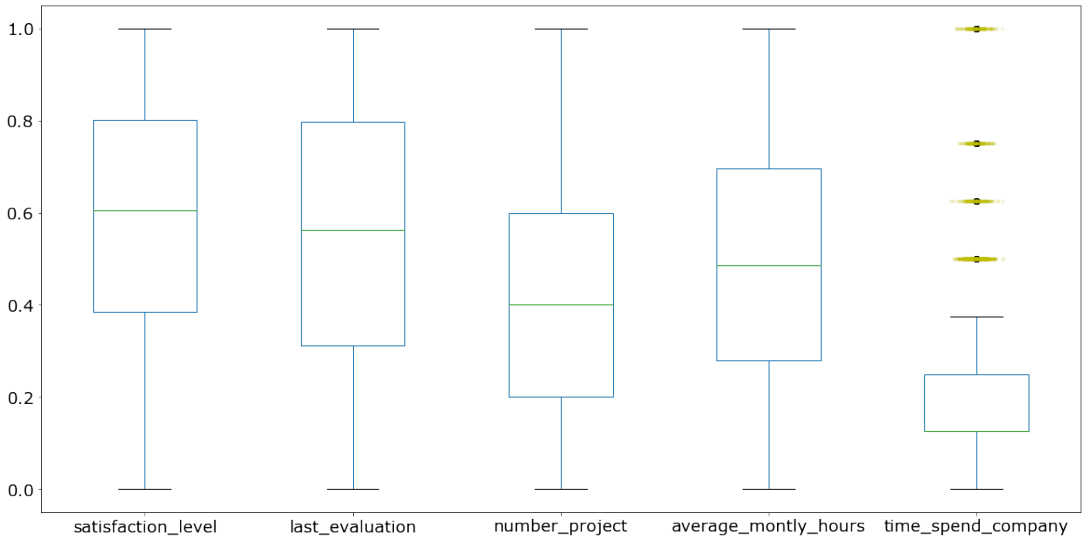


Figura 1.6: Rilevamento degli *outliers*

La figura 1.6 illustra la distribuzione dei valori degli attributi numerici: notiamo che `number_project` assume per lo più valori minori o uguali alla metà del valore massimo (in questo caso 7, vedere la tabella 1.2), mentre `time_spend_company` ha una distribuzione molto sbilanciata (il terzo quartile corrisponde al valore 4 e il valore massimo è 10, vedere ancora la tabella 1.2). In più essa è anche l'unica feature per la quale sono presenti degli *outlier*: nella figura 1.6 è stato aggiunto del *jitter* per avere un'idea della numerosità di tali valori. I suddetti *outlier* corrispondono ai valori la cui frequenza nell'istogramma, rispetto alla frequenza degli altri valori, è bassa (figura 1.2, istogramma relativo a `time_spend_company`).

Avendo appurato l'assenza di problemi nella qualità dei dati (vedere le assunzioni fatte all'inizio di questo paragrafo) e vista la quantità non trascurabile di *outlier*, è stato deciso di non escluderli dall'analisi.

1.4 Trasformazione degli attributi

A seconda delle necessità (formati di input richiesti da alcune librerie di supporto o semplicemente per rappresentare nella stessa scala diversi ordini di grandezza), talvolta i valori degli attributi numerici sono stati normalizzati nell'intervallo [0,1] e gli attributi categorici trasformati nei valori numerici corrispondenti.

In altri contesti, invece, alcuni attributi sono stati scartati da una specifica parte dell'analisi (`promotion_last_5years` nella parte dell'*Association Rules Mining*, per citarne uno) in quanto risultavano assai poco significativi.

Infine, nessuno degli attributi risulta essere legato ad altri da una stessa logica tale per cui sarebbe stato conveniente unirli in un unico attributo (per esempio tramite una funzione aritmetica che riassumesse più valori tramite una somma, differenza etc.), perciò nessuna trasformazione è stata effettuata in questa direzione.

1.5 Correlazione tra attributi ed eventuali variabili ridondanti

L'indice di correlazione di *Pearson* tra le coppie di attributi del dataset è rappresentato graficamente nella figura 1.7 (utilizzando una *Heat Map*). Nella maggior parte dei casi, il livello di correlazione non supera la soglia del 30% (si parla quindi di *correlazione debole*[1]), mentre se ci si concentra sulle feature `last_evaluation`, `number_project` e `average_montly_hours` notiamo che il livello di correlazione tra di esse si alza di poco. Tuttavia, esso rimane ben al di sotto della soglia della *correlazione forte* (che è del 70%). Date le circostanze, dunque, è stata esclusa la presenza di variabili ridondanti: di conseguenza, nessuna feature è stata esclusa dall'analisi.

[1] https://it.wikipedia.org/wiki/Indice_di_correlazione_di_Pearson

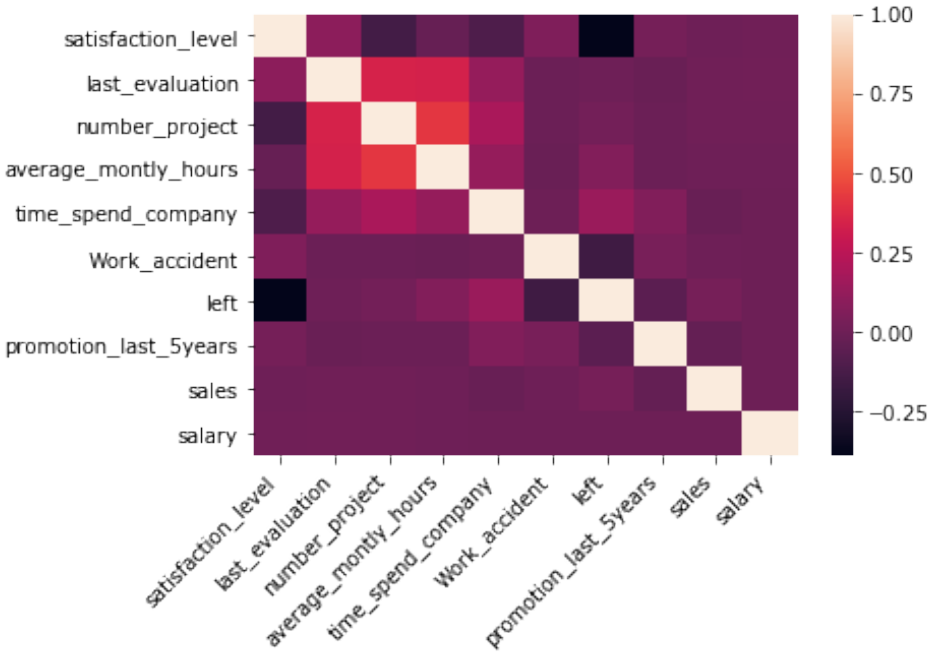


Figura 1.7: Correlazione tra gli attributi (metrica di *Pearson*)

2 Analisi dei cluster

L'analisi dei clusters è stata svolta normalizzando i valori degli attributi tra 0 e 1, con lo scopo di standardizzare il contributo del valore degli attributi nel calcolo della distanza.

2.1 clustering via K-means

2.1.1 Scelta degli attributi e della funzione distanza

Si sono scelti gli attributi quantitativi (*satisfaction_level*, *last_evaluation*, *average_montly_hours*, *time_spend_company*) e l'attributo numerico ordinale *number_project*. La scelta degli attributi quantitivi è giustificata dal fatto che l'algoritmo K-means richiede di calcolare la media, la quale è definita solo per gli attributi quantitiativi. L'unica eccezione è stata fatta per l'attributo *number_project* essendo oridinale. L'interpretazione è stata che se *number_project* ha valore frazionario $d.f$ (con d parte intera, f parte float) per un dato centroide, allora il cluster da lui rappresentato contiene gli impiegati che in media hanno fatto tra d e $d+1$ progetti.

La funzione distanza scelta è stata la distanza Euclidea perché i centroidi sono medie nello spazio Euclideo in $\Re^n$, dove n è il numero di attributi. Inoltre il valore della distanza tra 2 punti è di più facile interpretazione rispetto ad altre metriche.

2.1.2 Identificazione del miglior valore di k

Identificare il numero di clusters è importante per trovare un compromesso tra pochi grandi clusters e molti piccoli, e spesso insignificanti, clusters. Per indentificare il miglior valore di k per l'algoritmo K-means si è monitorato l'andamento della *Sum of Squared Error* (SSE) e la *silhouette score* al variare di k, il cui grafico è riportato in Figura 2.1. Il grafico della SSE ha un andamento non-crescente e diminuisce in maniera smooth, mentre la silhouette presenta molti picchi.

Sucessivamente si è scelto il punto di gomito della SSE combinando un approccio visivo ad uno quantitativo, analizzando i punti k in cui era presente un maggior calo di SSE. La scelta finale per il valore di k ha preso in considerazione anche il corrispondente valore della *silhouette*. La *silhouette* assumeva valori nel range $[0.18, 0.28]$ con minimo per $k = 47$ e massimo per $k = 3$, mentre la SSE assumeva valori tra $[640, 3315]$ con minimo per $k = 50$ e massimo per $k = 2$. Inoltre, sono stati presi in considerazione i top 10 valori di k corrispondenti a un maggiore calo della SSE (top_diffs). Analogamente sono stati presi i top 10 k con maggiore *silhouette* (top_silho). Si sono ottenuti i seguenti insiemi di valori di k, candidati ad essere punti di gomito: $top\_diffs = \{3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$ and $top\_silho = \{3, 8, 14, 7, 6, 10, 5, 9, 4, 18\}$. Mettendo insieme tutte le precendenti informazioni e considerando l'insieme $top\_diffs \cap top\_silho$, si è scelto il punto di gomito $k = 8$, corrispondente a $SSE = 1474$ and $silhouette = 0.274$.

Figura 2.1: Andamento della SSE e silhouette al variare di k. Il punto di gomito è scelto per $k = 8$, cui corrisponde $SSE = 1474$ e $silhouette = 0.27$.

Alternativamente, scegliendo il punto del grafico della SSE più vicino all'origine in norma Euclidea, si era ottenuto il punto $k = 12$, con $SSE = 1210$ e $silhouette = 0.257$.

2.1.3 Caratterizzazione dei clusters ottenuti

La caratterizzazione dei cluster ottenuti è stata svolta per mezzo dell'analisi dei centroidi, e confrontando le distribuzioni degli attributi dei singoli cluster con quelle dell'intero dataset.

La Figura 2.2 riporta l'analisi per centroidi. Gli attributi dei centroidi dei clusters 1 e 3 hanno relativamente lo stesso andamento. In particolare i valori di $average\_montly\_hours$ differiscono di poco. Questo significa che per valori più piccoli di k i due clusters potrebbero unirsi. I clusters 4, 5 e 7 sono caratterizzati da un basso valore di $satisfaction\_level$, mentre lo stesso attributo assume valori elevati nei clusters 0, 1, 3 e 6. Il cluster 6 è l'unico con un alto valore di $time\_spend\_company$, mentre per il resto dei clusters l'attributo assume valori bassi. Questo ci dice che i due attributi non sono correlati, altrimenti anche i clusters 0, 1 e 3 avrebbero avuto un alto valore di $time\_spend\_company$. Il cluster 4 ha il più basso valore di $satisfaction\_level$ e il più alto valore di $average\_montly\_hours$. Inoltre, il cluster 4 ha il più alto valore di $number\_projects$ e un relativamente basso valore di $time\_spend\_company$. Questo significa che il sovraccarico di lavoro dovuto ad una grande quantità di progetti, svolti in poco tempo, porta gli impiegati ad essere infelici.

In Figura 2.10 viene fatto un confronto tra le distribuzioni degli attributi dell'intero dataset e quelle dei singoli clusters, in particolare i cluster 4 e 7. L'attributo $number\_project$ assume una distribuizione multi-modale per tutte le kernel density estimation, essendo un attributo intero. Gli attributi $last\_evaluation$ e $average\_montly\_hours$ hanno una distribuzione simile alla normale, ma leggermente schiacciata. Gli stessi attributi assumono una simile forma per il cluster 4. Invece per l'intero dataset i due attributi hanno una distribuzione bi-modale con picchi poco pronunciati, divisi da un lungo plateau. In entrambe le figure Figura ?? e Figura ?? è presente un picco molto acuto per l'attributo $satisfaction\_level$. Entrambi i picchi sono quasi simmetrici, se non per lo scalino nella parte finale a destra del punto medio. Questo comportamento non è ripetuto per l'intero dataset. L'attributo $time\_spend\_company$ assume una distribuzione multi-modale per il cluster 7, con 5 picchi nell'intervallo $[0, 0.5]$. Mentre per il cluster 4, $time\_spend\_company$ ha meno variabilità riportando un picco relativamente acuto centrato in 0.25, e altri due picchi meno pronunciati.

2.1.4 Visualizzazione del clustering via Principal Component Analysis

I risultati ottenuti dal clustering via K-means sono stati combinati con la PCA per visualizzare la proiezione di ogni punto di dati in due dimensioni. Il dataset ha una forma globulare allungata, simile ad un ellisse. I punti di dati sono molto vicini gli uni agli altri formando un dataset molto denso, in cui i cluster 4 e 5 sono molto definiti, mentre i punti degli altri clusters tendono a mischiarsi. In basso a sinistra è presente una zona di bassa densità di punti appartenenti al cluster 0, mentre in alto a destra vi è un chiaro outlier assegnato al cluster 7. La natura dell'outlier in termini di attributi non è stata approfondita.

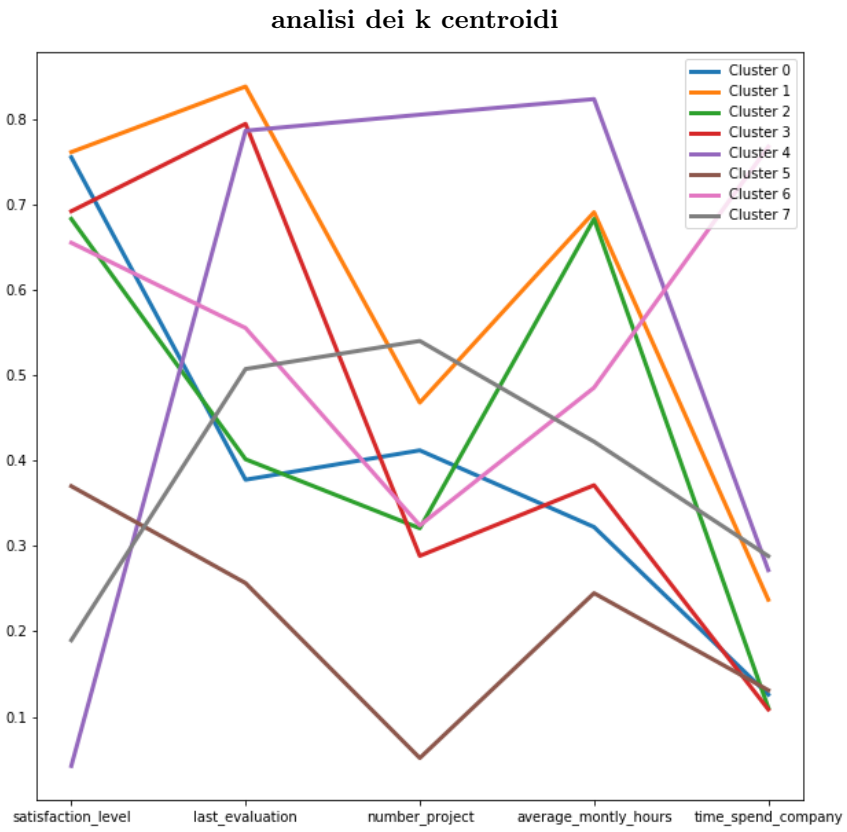


Figura 2.2: Analisi dei valori degli attributi dei centroidi ottenuti. Ogni centroide è rappresentato da un insieme di segmenti, i cui estremi marcano i valori degli attributi. Gli attributi dei clusters 1 e 3 seguono lo stesso andamento.

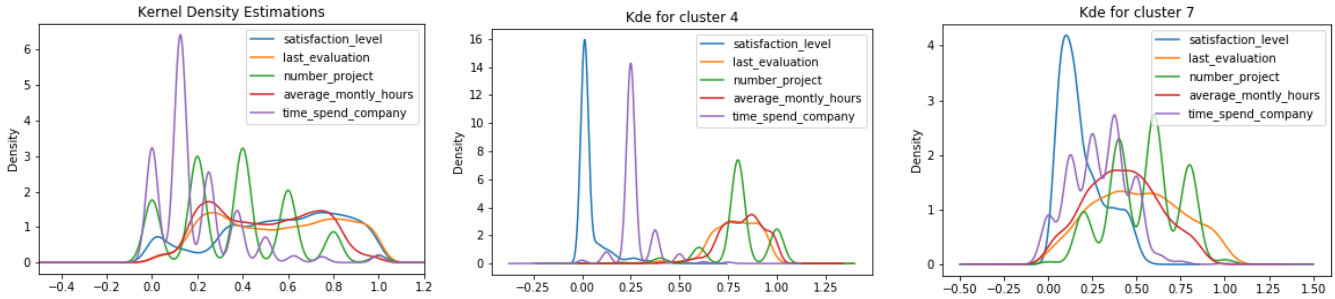


Figura 2.3: Kernel density estimation degli attributi per l'intero dataset, per il cluster 4, e per il cluster 7. Gli acuti picchi dell'attributo $satisfaction\_level$ per i cluster 4 e 7 non sono così pronunciati anche nell'intero dataset.

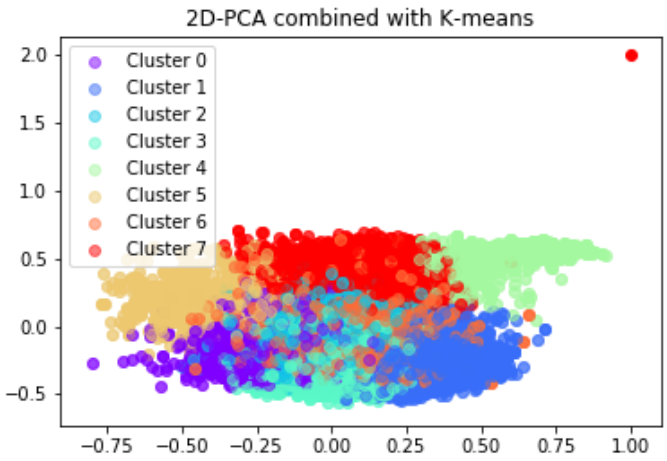


Figura 2.4: Visualizzazione del clustering in 2D. Ad ogni punto di dati è associata una label. Ogni label è rappresentata con un colore. Dataset denso a forma di ellisse. Un outlier è chiaramente presente in alto a destra.

2.2 Clustering via DBSCAN

L'analisi dei cluster con il metodo DBSCAN è stata svolta per identificare clusters con forma arbitraria. Con lo scopo di confrontare i risultati ottenuti con K-means, sono stati scelti gli attributi quantitativi e ordinali, e la distanza euclidea per funzione distanza.

2.2.1 Studio dei parametri minPoints ed epsilon

La scelta dei parametri $minPoints$ ed ϵ è stata ispirata dal metodo proposto da Ester et Al[2]. Figura 2.5 riporta il grafico delle distanze di ogni punto dati al $k-esimo$ punto dati più vicino, ordinate in modo decrescente. [3] I grafici, per valori diversi di k, avevano la stessa forma, e i valori delle ascisse erano uguali per almeno 7 punti decimali. Inoltre, dal momento che i valori di $minPoints$ ed ϵ giocano sullo stesso compromesso, l'analisi è stata portata avanti fissando $minPoints = k + 1$, escludendo il punto dati stesso dal calcolo dei punti vicini.

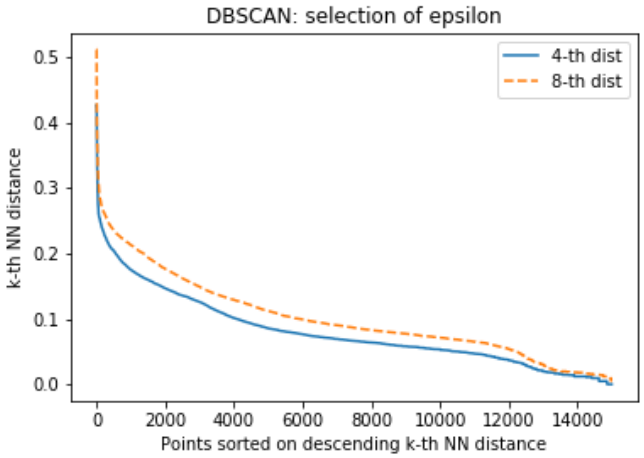


Figura 2.5: Distanze tra ogni punto dati e il suo $k-th$ punto più vicino ordinate in maniera non-crescente.

Per la scelta di ϵ, l'idea era quella di trovare il punto di gomito in Figura 2.5. Dalla figura si vede chiaramente che il punto di gomito giace nell'intervallo di valori $\epsilon \in [0.1, 0.3]$. Usando un metodo analogo utilizzato in sottosottosezione 2.1.2, si è ottenuto come risultato $\epsilon = 0.297$ e $silhouette = 0.301$. Dal momento che il valore ϵ ottenuto era ai margini dell'intervallo di valori predetto visualmente, è stata svolta un'indagine più approfondita per scoprire come variava il valore della $silhouette$ al variare di $\epsilon \in [0.1, 0.3]$. L'intervallo è stato diviso arbitrariamente in 30 punti. L'andamento della $silhouette$ è mostrato in Figura 2.6. Il grafico assume un plateau a partire da $\epsilon \cong 0.26$, in cui la $silhouette$ rimane quasi constante. Selezionando $\epsilon = 0.26$ si è ottenuto un insieme di due clusters rispettivamente di 14958 e 6 elementi, con valore di $silhouette = 0.344$. I rimanenti 35 punti dati sono stati classificati come noise-points.

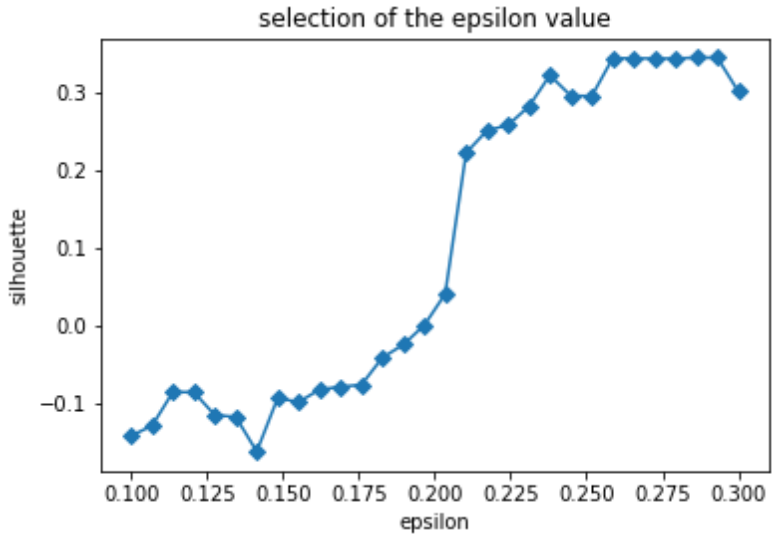


Figura 2.6: Andamento della silhouette per epsilon che varia nell'intervallo $[0.1, 0.3]$. Prima di $\epsilon = 0.2$ la silhouette è negativa, mentre è presente un grande incremento per $\epsilon \cong 0.225$. Dopo $\epsilon \cong 0.26$ è presente un plateau. Si è scelto $\epsilon = 0.26$, ottenendo $silhouette = 0.344$.

[2]Ester, Martin, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. Kdd. Vol. 96. No. 34. 1996.

[3]Per visibilità sono riportate soltanto le curve per $k = 4$ e $k = 8$. Le curve per valori intermedi di k giacciono tra le due curve mostrate in figura.

2.2.2 Caratterizzazione ed interpretazione dei clusters

I clustering ottenuti per diversi valori di ϵ sono stati caratterizzati analizzando la curva in Figura 2.6. Nell'intervallo di valori *epsilon* $\in [0.1, 0.2]$ (primi 15 punti sulla curva) il numero di clusters formati varia da 105 fino a 12, in cui sono presenti circa 6 clusters di medie dimensioni, e il resto di dimensioni molto piccole. I noise-points variano da 3278 fino a 281. In questo intervallo si ha *silhouette* ≤ 0, il che significa che mediamente i punti di un cluster sono più vicini ai punti di un altro cluster, piuttosto che ai punti del cluster di appartenenza, o che sono presenti troppi clusters. Un ulteriore interpretazione è che alcuni clusters si sovrappogono. Nella seconda metà del grafico, per $\epsilon \in (0.2, 0.3]$ si inizia a formare un unico grande cluster contenente almeno il 97% dei dati, con il rimanente 3% diviso tra pochi piccoli clusters e noise-points. In particolare, nell'intervallo $[0.26, 0.3)$ la *silhouette* forma un plateau, e sono presenti solo 2 clusters. Investigando i clusters ottenuti nei punti formanti il plateau si è scoperto che i noise-points gradualmente si uniscono ad uno dei 2 clusters. Infine, per $\epsilon = 0.3$ alcuni noise-points si uniscono per formare un piccolo cluster, diminuendo la *silhouette*.

2.3 Clustering Gerarchico

L'analisi dei cluster con approccio gerarchico è stata fatta con i metodi complete, average e di Ward, perché sono meno suscettibili al rumore dei dati e a gli outliers. Sono stati scelti gli attributi quantitativi e ordinali, e la distanza Euclidea, come in K-means e DBSCAN.[4]

Complete linkage. Figura ?? mostra il dendogramma ottenuto con il metodo complete linkage. A distanza minore di 25 si otterrebbero 16 clusters. I primi due clusters nella parte alta della figura (rispettivamente di 3300 e 4617 elementi) si uniscono con incremento di distanza relativamente alto, la quale diventa 1.6. Successive unificazioni di clusters incrementano lentamente la distanza fino a 1.91. Il dendogramma suggerisce la presenza di 2 clusters ottenuti con una linea di taglio a distanza 1.85.

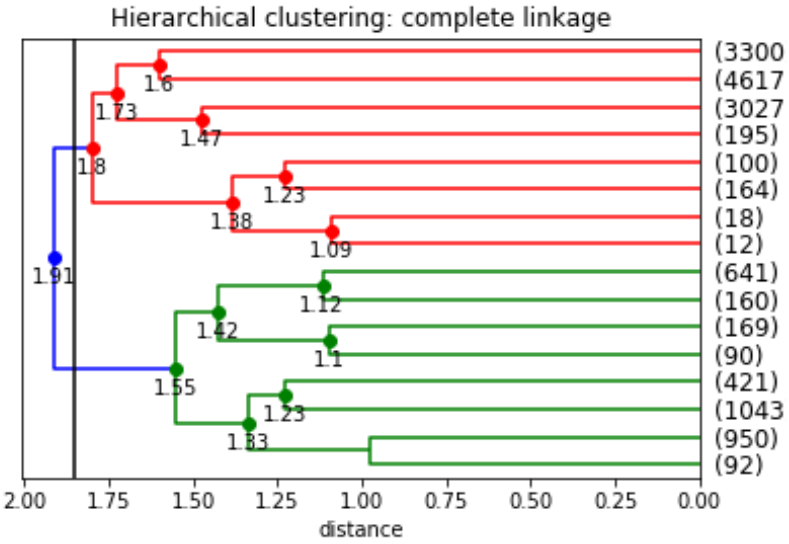


Figura 2.7: Dendogramma per il metodo complete linkage. La riga verticale nera rappresenta il taglio del dendogramma a distanza 1.85 risultante in 2 clusters.

Average Linkage. Figura ?? mostra il dendogramma ottenuto con il metodo average linkage. Nella parte bassa della figura, a distanza 1.04, un singleton (con label 7492) si unisce ad un grande agglomerato di clusters di 14966 elementi. Questa unificazione avviene relativamente tardi, il che suggerisce che il singleton sia un outlier. A metà figura due coppie di singleton si uniscono con un piccolo agglomerato di clusters, facendo un incremento di distanza relativamente alto, che arriva a 1.05. Dato questo grande salto in distanza e le dimensioni ridotte dei clusters coinvolti nell'unificazione, si è deciso di tagliare il dendogramma a distanza 1, ottenendo 4 clusters.

Ward's method Linkage. Figura ?? mostra il dendogramma ottenuto con il metodo di Ward. Il clustering ottenuto con questo metodo è il più equilibrato, in quanto nei livelli mostrati in figura, non sono presenti clusters di dimensioni molto piccole. Tuttavia la distanza assume valori molto più elevati rispetto ai precedenti metodi. Per una distanza minore di 15 si hanno unificazioni con incremento graduale della distanza, mentre negli ultimi 6 collegamenti la distanza fa lunghi salti raggiungendo 41.61. Per questo motivo si è deciso di tagliare il dendogramma a distanza 15 ottenendo 7 clusters.

[4]Sono visualizzati soltanto i primi 3 livelli dei dendogrammi per facilitarne la comprensione.

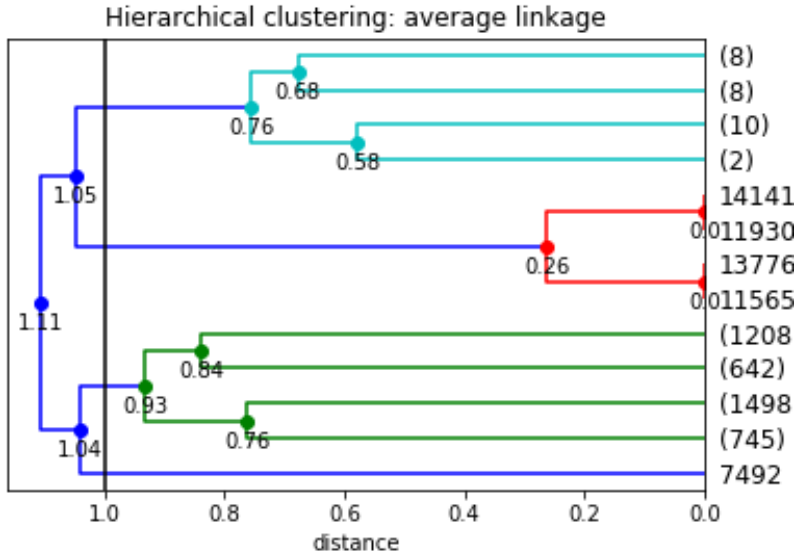


Figura 2.8: Dendogramma per il metodo average linkage. La riga che taglia il dendogramma a distanza 1 risulta in 4 clusters.

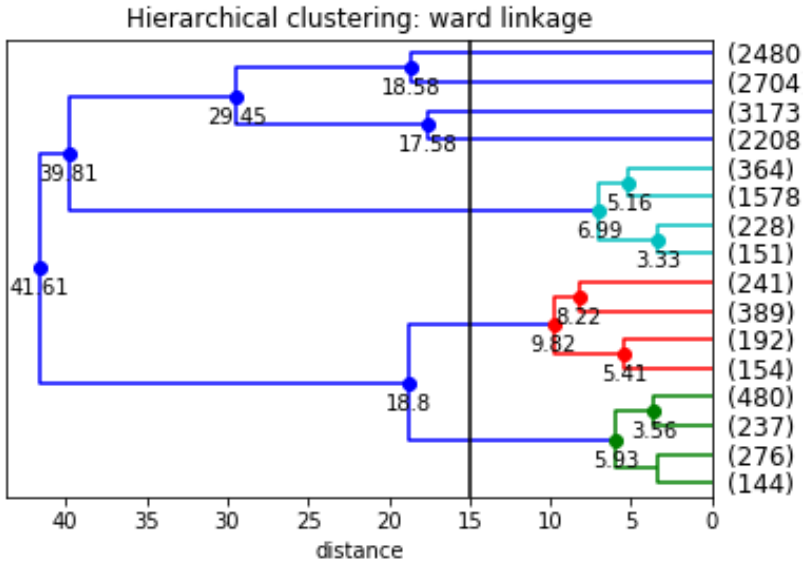


Figura 2.9: Dendogramma per il metodo Ward. Il taglio a distanza 15 fa ottenere 7 clusters.

Tabella 2.1: Intervallo di valori assunti dalla silhouette, per un numero di cluster C da 2 a 9, per i metodi complete, average e Ward. Risultati ottenuti con un algoritmo non strutturato (U) e strutturato considerando n=100 vicini. Per ogni metodo il miglior risultato è mostrato in grassetto. I numeri sottolineati sono risultati altrettanto buoni.

Metodo	Silhouette C=2 - U	Silhouette C=9 - U	Silhouette C=2 - n=100	Silhouette C=9 - n=100
Complete	0.197	0.143	**0.348**	- 0.061
Average	0.339	0.209	**0.348**	0.149
Ward	**0.306**	0.215	0.295	0.202

Tabella 2.2: Confronto delle prestazioni degli algoritmi di clustering K-means, DBSCAN e gerarchico. Risultati ottenuti impostando i seguenti parametri: K=8 per K-means. MinPoints=4, $\epsilon = 0.26$ per DBSCAN ottenendo 2 clusters. Metodo average linkage con n=100 vicini per clustering gerarchico, ottenendo 2 clusters. In grassetto il miglior risultato.

Algoritmo	Silhouette
K-means	0.274
DBSCAN	0.344
Gerarchico	**0.348**

Calcolo della silhouette. Il calcolo della silhouette è stato fatto considerando un algoritmo di clustering non strutturato, e uno di tipo strutturato considerando $n = 100$ punti vicini[5]. La silhouette è stata calcolata per i tre metodi impostando un numero di clusters da 2 a 9. Il range di valori assunti dalla silhouette è mostrato in Tabella 2.1, in cui si vede che i migliori risultati si ottengono per C=2 clusters e n=100 per i metodi complete e average, e con algoritmo non strutturato per il metodo Ward. In particolare, con i primi due metodi si ottengono risultati identici in termini di silhouette (solo per C=2). Anche i clusters ottenuti hanno le stesse dimensioni (solo per C=2) di 14995 e 4 elementi[6].

2.4 Valutazione del miglior metodo di clustering

Tabella 2.2 riassume le prestazioni degli algoritmi di clustering usati. L'algoritmo che ha dato risultati migliori, in termini di silhouette, è il clustering gerarchico usando il metodo average linkage[7] , il quale rileva che nel dataset sono presenti 2 clusters, rispettivamente di 14955 e 4 elementi. DBSCAN ha avuto risultati simili, sia in termini di silhouette che per i clusters rilevati, ottenendo 2 clusters di dimensioni 14958 e 6 elementi, e 35 noise-points. Mentre l'algoritmo con prestazioni minori è K-means, il quale rileva 7 clusters di dimensioni quasi omogenee.

3 Pattern e Association Rules mining

In questa sezione si presenta il processo di analisi delle *Association Rules*. Inizialmente, si discutono alcune azioni preliminari effettuate sul dataset per preparare i dati alle operazioni successive. Dopodiché, si attua l'estrazione degli itemset frequenti (*massimali*, *chiusi* e *frequenti*). Infine si estraggono da tali itemset le più interessanti regole di associazione, anche con lo scopo di costruire un modello di predizione per i valori mancanti e per l'attributo `left`.

3.1 Operazioni preliminari

Una delle operazioni propedeutiche alla fase di mining multidimensionale delle regole di associazione è la discretizzazione degli attributi numerici. Questa è stata effettuata in parte sulla base delle distribuzioni degli attributi in questione (quando ritenute interessanti) e in parte definendo intervalli di ampiezza fissata.

I valori dell'attributo `satisfaction_level` sono stati raggruppati in quattro classi: *very_ low*, per valori nell'intervallo [0, 0.25); *low*, per valori nell'intervallo [0.25, 0.50); *medium*, per valori nell'intervallo [0.50, 0.75); *high*, per valori nell'intervallo [0.75, 1]. Tale suddivisione è stata ispirata dalla distribuzione lievemente irregolare dell'attributo (si veda la Figura 3.1), che evidenzia un piccolo pinnacolo tra 0 e 0.25.

L'attributo `last_evaluation` è stato invece discretizzato utilizzando intervalli di ampiezza 0.2, alla luce di una distribuzione priva di irregolarità ritenute significative.

Lo stesso vale per `average_montly_hours`, per il quale sono stati scelti intervalli di ampiezza 30. L'idea per la scelta di tale ampiezza è che un'ora (in media) di lavoro al giorno di differenza sia una misura di discriminazione adeguata.

Dal momento che l'attributo `number_project` presenta una distribuzione pressoché uniforme, si è scelto di discretizzarlo con intervalli di ampiezza fissata, in particolare 2, ritenendo che un solo progetto di differenza fosse una misura di discriminazione troppo "a grana fine", e che pertanto potesse dare origine a pattern frequenti e regole di associazione ridondanti.

[5] Dato il costo computazionale, valori maggiori di n non sono stati considerati.
[6] Non è stato controllato che anche i singoli elementi fossero assegnati a gli stessi clusters.
[7] A pari merito con il metodo complete linkage.

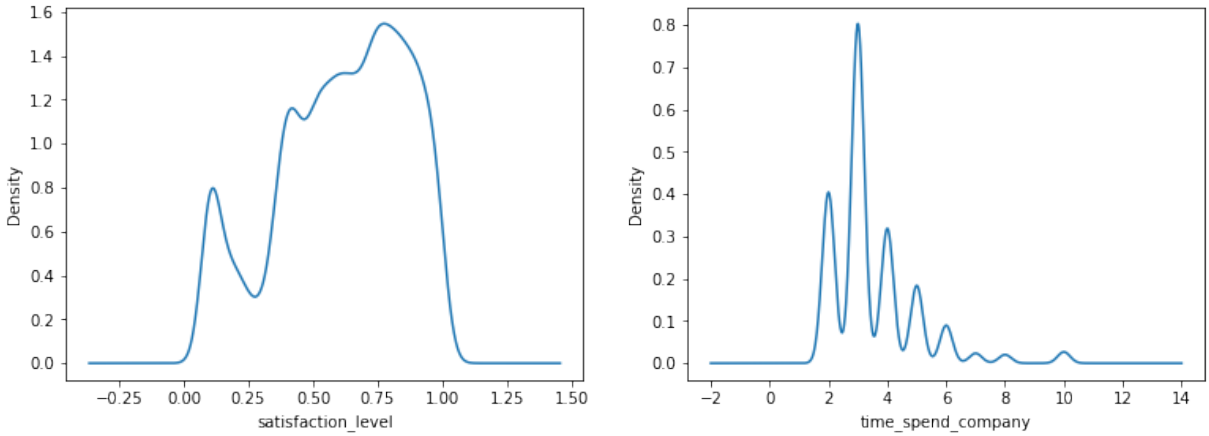


Figura 3.1: Kernel density estimation degli attributi satisfaction _level e time_spend_company.

Per la discretizzazione di `time_spend_company`, infine, la scelta dell'ampiezza degli intervalli riflette l'irregolarità della distribuzione (si veda la Figura 3.1). Gli intervalli scelti, pensati appositamente per identificare tre categorie ben distinte di impiegati (che potremmo caratterizzare rispettivamente come quella degli impiegati recentemente assunti, quella di chi ha già alcuni anni alle spalle ed infine quella dei veterani), sono [2, 4), [4, 7) e [7, 10].

Per l'interpretazione dei valori discretizzati si faccia riferimento alla Tabella 3.1. I valori dell'attributo `sales` non hanno alcun suffisso in quanto identificabili senza alcuna ambiguità, mentre i valori per gli attributi `number_project`, `average_montly_hours` e `last_evaluation` sono da intendersi come gli estremi sinistri del corrispondente intervallo di appartenenza (nel caso di `last_evaluation` il valore rappresenta una percentuale, perciò 50_LE, ad esempio, indica in realtà un valore nell'intervallo [0.5, 0.7)).

Prima di procedere con l'estrazione degli itemset frequenti e delle regole di associazione, si è deciso inoltre di rimuovere interamente l'attributo `promotion_last_5years`. La ragione di questa scelta è che oltre il 97% delle righe del dataset hanno il medesimo valore dell'attributo, ossia 0. Ciò significa che la stragrande maggioranza degli itemset frequenti e delle regole di associazione registrerebbero il valore 0 per `promotion_last_5years`, il che appesantirebbe soltanto l'interpretazione degli itemset e delle regole senza rivelare alcuna proprietà interessante.

Suffisso	Attributo corrispondente
_WA	Work_accident
_L	left
_SAT	satisfaction_level
_SAL	salary
_LE	last_evaluation
_AMH	average_montly_hours
_NP	number_project
_TSC	time_spend_company

Tabella 3.1: Legenda dei valori discretizzati

3.2 Estrazione degli itemset frequenti

Le sezioni che seguono sono dedicate all'estrazione delle diverse tipologie di itemset (massimali, chiusi e frequenti) con diversi valori del supporto. Nel caso degli itemset frequenti e chiusi si è scelto di restringere lo spazio di ricerca a quelli contenenti almeno tre elementi, mentre per quelli massimali il vincolo è di almeno due elementi. Per ogni intervallo di valori del supporto, i 10 itemset (quando disponibili) ritenuti più significativi sono stati riportati in una tabella. Di questi sono stati commentati i più interessanti.

3.2.1 Itemset massimali

Nella Tabella 3.2 sono riportati gli itemset massimali più significativi, estratti restringendo il valore del supporto tra 15 e 20 (in percentuale). Gli itemset 2 e 3 rivelano che una piccola percentuale degli impiegati che hanno trascorso pochi anni in azienda (da 2 a 3) - ed attualmente con un qualche incarico all'interno di essa - hanno un valore della valutazione tra 0.5 e 0.9. L'itemset 1 rivela, senza sorprese, che simili impiegati hanno concluso un numero di progetti

relativamente basso (da 2 a 4). Il fatto che tutti gli itemset evidenzino l'assenza di incidenti sul lavoro è probabilmente imputabile alla distribuzione non omogenea dell'attributo `Work_accident`, che assume il valore 0 per più dell'80% degli impiegati (si veda la Figura 1.1 e la Tabella 1.2).

#	Itemset	Supporto (%)
1	{2_NP, 2_to_3_TSC, 0_L, 0_WA}	19.5946
2	{50_LE, 2_to_3_TSC, 0_L, 0_WA}	17.7945
3	{70_LE, 2_to_3_TSC, 0_L, 0_WA}	15.5677
4	{4_to_6_TSC, 0_L, 0_WA}	15.4277

Tabella 3.2: Itemset massimali con supporto tra 15% e 20%

Nella Tabella 3.3 sono invece registrati gli itemset massimali con supporto tra 10 e 15. Tra i più interessanti troviamo i numeri 5 e 6, che rispettivamente registrano, per alcuni degli impiegati con un salario basso, l'abbandono dell'impiego (come è lecito aspettarsi) ed un valore alto dell'ultima valutazione (rivelando un dato insapettato), e il numero 7, il quale suggerisce che, per una percentuale non trascurabile di impiegati, un incidente sul lavoro non influisce sulla di questi volontà di mantenere il posto in azienda. La considerazione fatta in precendenza circa l'assidua presenza di 0_WA negli itemset trova in questo caso un ulteriore riscontro.

#	Itemset	Supporto (%)
1	{2_NP, low_SAL, 2_to_3_TSC, 0_WA}	14.8077
2	{4_to_6_TSC, 4_NP, 0_WA}	14.541
3	{4_to_6_TSC, low_SAL, 0_WA}	14.2543
4	{50_LE, 2_NP, 2_to_3_TSC, 0_WA}	14.1476
5	{1_L, low_SAL, 0_WA}	13.8476
6	{70_LE, low_SAL, 0_WA}	13.6876
7	{1_WA, 0_L}	13.3342
8	{150_AMH, 0_L, 0_WA}	13.2942
9	{150_AMH, 2_to_3_TSC, 0_WA}	13.0475
10	{50_LE, 4_NP, 0_L, 0_WA}	12.6808

Tabella 3.3: Itemset massimali con supporto tra 10% e 15%

Itemset massimali con supporto maggiore o uguale a 20 non sono stati rilevati.

3.2.2 Itemset chiusi

Gli unici due itemset chiusi con supporto superiore al 30% sono contenuti nella Tabella 3.4. Il numero 1 rivela che una buona frazione degli impiegati (circa il 44 %) è costituita da coloro che hanno trascorso poco tempo in azienda, non sono stati vittime di incidenti e mantengono tuttora il loro incarico. Essi potrebbero rappresentare i tipici nuovi arrivati. Il numero 2 cattura un altrettanto buona porzione di impiegati aventi caratteristiche simili (nessun incidente e nessun abbandono del lavoro) ed un numero di progetti portati a termine tra 4 e 5.

#	Itemset	Supporto (%)
1	{2_to_3_TSC, 0_L, 0_WA}	44.4696
2	{4_NP, 0_L, 0_WA}	33.7489

Tabella 3.4: Itemset chiusi con supporto maggiore del 30%

Gli itemset chiusi con supporto tra 20% e 30% possono essere osservati nella Tabella 3.5. Similmente ad alcuni itemset riportati e discussi precedentemente, il numero 1 rivela caratteristiche non troppo sorprendenti di quelli che sembrano essere gli impiegati con un trascorso breve all'interno dell'azienda: un basso numero di progetti realizzati (2 o 3), pochi anni di esperienza (2 o 3) e nessun incidente lavorativo. Il numeri 3 e 6 sembrano invece identificare due prototipi dell'impiegato standard, da una parte accomunati dall'assenza di incidenti e dalla permanenza in azienda, dall'altra contraddistinti rispettivamente da un salario medio e da un livello medio di soddisfazione. I numeri 4 e 5 rappresentano due categorie simili di impiegati, concordanti sul salario basso e sull'assenza di incidenti lavorativi, ma con le rispettive peculiarità di non aver abbandonato il lavoro (nonostante il salario basso) e di aver passato da 2 a 3 anni in azienda (che potrebbe suggerire che salari medi ed alti siano riservati a chi ha più esperienza).

3.2.3 Itemset frequenti

La Tabella 3.6 mostra gli itemset frequenti con supporto superiore al 20%. Il fatto che nessuno di questi abbia un supporto maggiore o uguale al 25% può sembrare contraddittorio, visto che gli itemset frequenti sono un sovrainsieme degli itemset massimali e chiusi elencati in precedenza. In realtà, ciò è dovuto al fatto che sono pochi gli itemset

#	Itemset	Supporto (%)
1	{2_NP, 2_to_3_TSC, 0_WA}	29.4753
2	{4_NP, 2_to_3_TSC, 0_L}	28.6886
3	{medium_SAL, 0_L, 0_WA}	28.4419
4	{low_SAL, 0_L, 0_WA}	27.9952
5	{low_SAL, 2_to_3_TSC, 0_WA}	26.8418
6	{medium_SAT, 0_L, 0_WA}	26.4284
7	{2_NP, 0_L, 0_WA}	26.3084
8	{low_SAL, 2_to_3_TSC, 0_L}	24.8283
9	{high_SAT, 0_L, 0_WA}	24.6616
10	{medium_SAT, 2_to_3_TSC, 0_L}	24.2683

Tabella 3.5: Itemset chiusi con supporto tra 20% e 30%

propriamente frequenti, ossia non chiusi. Per questo, gli itemset frequenti con supporto maggiore del 25% che sembrano "mancanti" sono in realtà chiusi e, in quanto tali, non sono stati riportati una seconda volta sotto la sezione degli itemset frequenti.

Cercando di individuare gli itemset più significativi, il numero 1 testimonia che una buona percentuale di impiegati caratterizzati da una breve permanenza in azienda (2 o 3 anni) sono riusciti a portare a termine un discreto numero di progetti (4 o 5) senza incorrere in incidenti. Ciò fa pensare a progetti relativamente semplici ed esenti da rischi. Il numero 6 rivela invece l'esistenza di alcuni impiegati che, malgrado i pochi anni trascorsi nella compagnia, sono riusciti ad ottenere un salario medio e compatibilmente scelgono di mantenere il posto di lavoro. Una categoria analoga di persone è identificata dall'itemset numero 8, con l'unica differenza che, al posto di un salario medio, questa include coloro che possono ritenersi soddisfatti del loro impiego (alto livello di soddisfazione).

#	Itemset	Supporto (%)
1	{4_NP, 2_to_3_TSC, 0_WA}	24.2283
2	{50_LE, 2_to_3_TSC, 0_WA}	24.1216
3	{50_LE, 0_L, 0_WA}	23.9416
4	{4_NP, 2_to_3_TSC, 0_L, 0_WA}	23.8416
5	{2_NP, 2_to_3_TSC, 0_L}	23.7349
6	{medium_SAL, 2_to_3_TSC, 0_L}	23.6482
7	{medium_SAL, 2_to_3_TSC, 0_WA}	23.4816
8	{high_SAT, 2_to_3_TSC, 0_L}	22.6215
9	{70_LE, 0_L, 0_WA}	22.3882
10	{50_LE, 2_to_3_TSC, 0_L}	21.3014

Tabella 3.6: Itemset frequenti con supporto maggiore del 20%

La Tabella 3.7 registra invece gli itemset frequenti con valore del supporto tra 10% e 20%. Il numero 4 mette in luce un gruppo di impiegati che sono riusciti a soddisfare le aspettative dell'azienda, totalizzando un punteggio tra 0.7 e 0.9 nell'ultima valutazione, nonostante un trascorso breve. Senza sorprese, tali impiegati scelgono di non lasciare il lavoro. Questa categoria di dipendenti potrebbe rappresentare coloro che sono partiti "col piede giusto". Il numero 6 cattura invece un insieme di persone aventi un alto livello di soddisfazione ed numero di progetti completati compreso tra 4 e 5, il tutto accompagnato dall'assenza di incidenti sul posto di lavoro. Per questi impiegati è possibile concludere che soddisfazione e produttività vanno di pari passo.

#	Itemset	Supporto (%)
1	{medium_SAL, 2_to_3_TSC, 0_L, 0_WA}	19.7813
2	{4_NP, low_SAL, 0_WA}	19.1813
3	{high_SAT, 2_to_3_TSC, 0_WA}	18.9879
4	{70_LE, 2_to_3_TSC, 0_L}	18.9613
5	{high_SAT, 2_to_3_TSC, 0_L, 0_WA}	18.8346
6	{high_SAT, 4_NP, 0_WA}	18.8013
7	{4_NP, low_SAL, 0_L}	18.5612
8	{medium_SAL, 4_NP, 0_L}	18.1679
9	{2_NP, low_SAL, 0_WA}	17.9812
10	{medium_SAL, 4_NP, 0_WA}	17.5612

Tabella 3.7: Itemset frequenti con supporto tra 10% e 20%

3.3 Estrazione delle regole di associazione

Di seguito si riportano le regole di associazione estratte per differenti valori della confidenza. Al fine di limitare il numero di regole restituite, si è scelto di generarle a partire da itemset frequenti di lunghezza maggiore o uguale a 3. Per ogni intervallo di valori della confidenza, le 10 regole ritenute più significative sono state riportate in una tabella, quindi si è proceduto a commentare quelle più interessanti. In ciascuna tabella è possibile distinguere tre categorie di regole:

- regole "generiche", la cui conseguenza rivela un dato riguardante un attributo diverso da `left`
- regole la cui conseguenza rivela l'abbandono del posto di lavoro (`left`=1)
- regole la cui conseguenza rivela il mantenimento del posto di lavoro (`left`=0)

La Tabella 3.8 riporta le regole con confidenza maggiore del 90%. La numero 3 indica che poche ore di lavoro (da 120 a 150 al mese), combinate con un piccolo numero di progetti completati (2 o 3), una breve permanenza in azienda e l'abbandono dell'impiego, determinano quasi certamente (con una confidenza maggiore del 99%) un basso livello di soddisfazione. La numero 4 varia la causa sostituendo le poche ore di lavoro con un valore dell'ultima valutazione tra 0.5 e 0.7, rivelando la medesima conseguenza. La numero 8 evidenzia una delle principali combinazioni di fattori che portano all'abbandono del posto di lavoro, ossia un gran numero di progetti completati, una permanenza in azienda compresa tra 4 e 6 anni, nonché un livello di soddisfazione molto basso. A questi va aggiunto anche il non verificarsi di incidenti, come a sottolineare che non si tratta del vero motivo per cui gli impiegati decidono di andarsene. Questa regola suggerisce che nel lungo termine i dipendenti sono più propensi a lasciare il lavoro. La numero 9, invece, indica che un alto livello di soddisfazione tra quelli che potremmo definire i neo assunti (ossia che hanno trascorso dai 2 ai 3 anni in azienda) fa sì che questi scelgano di rimanere.

#	**Premessa**	**Conseguenza**	**Lift (%)**	**Confidenza (%)**
1	{very_low_SAT, 1_L, 4_to_6_TSC}	6_NP	982.029	93.6264
2	{6_NP, 1_L, 4_to_6_TSC}	very_low_SAT	829.409	96.1625
3	{120_AMH, 1_L, 2_NP, 2_to_3_TSC}	low_SAT	526.352	99.803
4	{1_L, 50_LE, 2_NP, 2_to_3_TSC}	low_SAT	522.447	99.0625
5	{6_NP, very_low_SAT, 1_L}	4_to_6_TSC	312.963	99.0698
6	{120_AMH, low_SAT, 1_L, 2_to_3_TSC}	2_NP	231.198	99.3137
7	{120_AMH, low_SAT, 2_NP, 2_to_3_TSC}	1_L	383.663	91.3436
8	{6_NP, very_low_SAT, 4_to_6_TSC, 0_WA}	1_L	379.931	90.455
9	{high_SAT, 2_to_3_TSC}	0_L	130.326	99.2976
10	{4_NP, 2_to_3_TSC}	0_L	129.473	98.6474

Tabella 3.8: Regole di associazione con confidenza maggiore del 90%

La Tabella 3.9 riporta le regole con confidenza tra l'80% e il 90%. La regola 1 mette in risalto un dato molto interessante: un numero di progetti tra 6 e 7, accompagnato da una permanenza relativamente lunga (da 4 a 6 anni), comporta un livello di soddisfazione molto basso. Ciò rivela un'informazione preziosa per l'azienda: i dipendenti con più esperienza non si ritengono affatto soddisfatti della loro attuale situazione, pertanto è bene correre ai ripari. La numero 4 suggerisce che le cause di una scarsa efficienza (solo 2 o 3 progetti completati) sono un livello di soddisfazione basso ed un punteggio medio basso (da 0.3 a 0.5) nell'ultima valutazione. È probabile quindi che una migliore produttività si possa ottenere a fronte di incentivi da parte della compagnia, siano essi diretti (fornire un punteggio maggiore nella valutazione) o indiretti (cercare di aumentare il livello di soddisfazione dei lavoratori). La regola 7 mostra che un numero di ore mensili compreso tra 120 e 150, un numero di progetti realizzati compreso tra 2 e 3 ed un basso livello di soddisfazione causano nell'88% dei casi l'abbandono del posto di lavoro. Contrariamente, la regola 10 afferma che un livello di soddisfazione alto ed un contesto lavorativo sicuro sono fattori che determinano il mantenimento del posto di lavoro.

#	**Premessa**	**Conseguenza**	**Lift (%)**	**Confidenza (%)**
1	{6_NP, 4_to_6_TSC}	very_low_SAT	704.962	81.734
2	{120_AMH, 2_NP, low_SAL, 2_to_3_TSC, 0_WA}	low_SAT	425.47	80.6744
3	{low_SAT, 50_LE, low_SAL}	2_NP	201.126	86.3962
4	{30_LE, low_SAT}	2_NP	197.358	84.7775
5	{sales, 2_NP, low_SAL}	2_to_3_TSC	126.594	81.7597
6	{4_to_6_TSC, high_SAT, 4_NP}	0_WA	105.14	89.9356
7	{120_AMH, low_SAT, 2_NP}	1_L	371.404	88.4247
8	{6_NP, very_low_SAT, 4_to_6_TSC}	1_L	368.547	87.7446
9	{medium_SAL, 4_NP}	0_L	115.037	87.6488
10	{high_SAT, 0_WA}	0_L	108.829	82.9186

Tabella 3.9: Regole di associazione con confidenza tra 80% e 90%

Nella Tabella 3.10 sono elencate le regole con confidenza compresa tra 70% ed 80%. La regola 5 rivela che gli impiegati con un trascorso piuttosto lungo (da 4 a 6 anni), un numero di progetti alle spalle compreso tra 4 e 5, un punteggio molto alto nell'ultima valutazione (tra 0.9 e 1) e nessun incidente si ritengono, nella maggior parte dei casi (circa 70%), molto soddisfatti. Questo suggerisce che i dipendenti prediligono un carico di lavoro relativamente basso (circa 1 progetto all'anno), il quale consente loro, peraltro, di ottenere un'ottima valutazione da parte dell'azienda. La regola 7 mette invece in evidenza il fatto che un salario ed un livello di soddisfazione bassi, abbinati ad una produttività mediocre (2 o 3 progetti completati), determina la perdita dell'impiego. La regola 10, infine, dimostra che gli impiegati assunti da poco tempo sono tendenzialmente disposti ad accettare un salario basso pur di mantenere il loro posto in azienda.

#	**Premessa**	**Conseguenza**	**Lift (%)**	**Confidenza (%)**
1	{very_low_SAT, 4_to_6_TSC, 70_LE}	6_NP	764.086	72.8477
2	{6_NP, low_SAL}	very_low_SAT	663.809	76.9627
3	{30_LE, 2_NP, 2_to_3_TSC}	low_SAT	409.318	77.6119
4	{1_L, 4_to_6_TSC, 4_NP}	high_SAT	223.705	78.7942
5	{90_LE, 4_to_6_TSC, 4_NP, 0_WA}	high_SAT	199.63	70.3145
6	{low_SAT, low_SAL}	2_NP	185.483	79.6764
7	{low_SAT, 2_NP, low_SAL}	1_L	324.144	77.173
8	{6_NP, 4_to_6_TSC}	1_L	313.249	74.5791
9	{70_LE, 0_WA}	0_L	103.409	78.7893
10	{low_SAL, 2_to_3_TSC}	0_L	103.29	78.6982

Tabella 3.10: Regole di associazione con confidenza tra 70% e 80%

3.3.1 Predizione dei valori mancanti

In assenza di valori mancanti all'interno del dataset, si è scelto di introdurne appositamente alcuni. In particolare, si è tentato di predire, utilizzando le regole più significative, i valori dell'attributo `Work_accident` per una frazione del dataset originale. A questo scopo, si è adottato il seguente procedimento: per prima cosa è stato estratto in maniera casuale il 10% delle righe (mantenendo la distribuzione relativa dei valori dell'attributo `Work_accident`). Fatto ciò, sono state estratte le regole di associazione aventi rispettivamente `0_WA` e `1_WA` nella conseguenza, restringendo il campo a quelle con valore della confidenza maggiore o uguale al 70%. Per ottenere l'insieme di regole riportate nella Tabella 3.11 sono state estratte le prime 5 regole (ordinate secondo il valore del lift) di ciascuno dei due insiemi. Si noti che nella tabella non appare nessuna regola avente `1_WA` come valore della conseguenza, questo perché nessuna regola con tali caratteristiche è stata restituita da Apriori. La predizione dei valori mancanti con tale insieme di regole raggiunge un'accuratezza del 98.4%. Qualora si desiderasse ripetere l'esperimento, è sufficiente estrarre casualmente il 10% delle righe del dataset (avendo l'accortezza di preservare la distribuzione relativa dei valori di `Work_accident`) utilizzando la funzione Python `pandas.DataFrame.sample` e impostando il valore del parametro `random_state` a 1.

#	**Premessa**	**Conseguenza**	**Lift (%)**	**Confidenza (%)**
1	{120_AMH, low_SAT, 1_L}	0_WA	112.284	96.0463
2	{90_LE, 1_L, 4_to_6_TSC}	0_WA	112.265	96.03
3	{90_LE, 1_L}	0_WA	112.259	96.0251
4	{120_AMH, low_SAT, 1_L, 2_to_3_TSC}	0_WA	112.207	95.9804
5	{120_AMH, low_SAT, 1_L, 2_NP}	0_WA	112.188	95.9646

Tabella 3.11: Regole di associazione usate per predire il valore di Work_accident

3.3.2 Predizione dell'attributo 'left'

Per predire i valori dell'attributo `left` si è proceduto in maniera del tutto analoga. Regole aventi il valore `1_L` e `0_L` nella conseguenza sono state separatamente estratte ed ordinate secondo il valore del lift. Da ciascuno dei due insiemi ordinati sono state estratte le prime 5 regole, le quali sono state combinate per ottenere l'insieme mostrato nella Tabella 3.12. L'impiego di tali regole per la predizione dei valori di `left` (attenendosi al verdetto della maggioranza nel caso di più regole compatibili con una stessa riga del dataset) permette di raggiungere un'accuratezza del 96.7%.

4 Classificazione

In questa sezione si illustra il processo di classificazione. L'obiettivo è quello di costruire un modello basato su *Decision tree* e utilizzarlo per prevedere se, a partire dai valori di alcuni suoi attributi, un impiegato dell'azienda lascerà o meno il lavoro. Nello specifico, gli attributi utilizzati in questa sede sono `satisfaction_level,last_evaluation,number_project,average_montly_hours,time_spend_company,Work_accident,salary`. Tra questi

#	Premessa	Conseguenza	Lift (%)	Confidenza (%)
1	{120_AMH, low_SAT, 2_NP, 2_to_3_TSC, 0_WA}	1_L	388.451	92.4833
2	{120_AMH, low_SAT, 2_NP, 2_to_3_TSC}	1_L	383.663	91.3436
3	{6_NP, very_low_SAT, 4_to_6_TSC, 0_WA}	1_L	379.931	90.455
4	{120_AMH, low_SAT, 2_NP, 0_WA}	1_L	377.439	89.8618
5	{120_AMH, low_SAT, 2_to_3_TSC, 0_WA}	1_L	375.184	89.3248
6	{high_SAT, 2_to_3_TSC}	0_L	130.326	99.2976
7	{medium_SAT, 2_to_3_TSC}	0_L	129.786	98.8862
8	{medium_SAT, 2_to_3_TSC, 0_WA}	0_L	129.521	98.6842
9	{4_NP, 2_to_3_TSC}	0_L	129.473	98.6474
10	{4_NP, 2_to_3_TSC, 0_WA}	0_L	129.153	98.404

Tabella 3.12: Regole di associazione usate per predire il valore di left

attributi, quelli categorici sono stati codificati con valori interi. É stato deciso di non prendere in considerazione l'attributo categorico non ordinale `sales` perché il suo valore risultava, nella maggior parte dei casi, totalmente ininfluente nel processo di decisione. L'importanza di tale feature nei processi di decisione dei modelli 1, 2, 3, 4 (illustrati nel seguito), infatti, è di 0, 0, 0 e 0.005 rispettivamente.

4.1 Classificazione tramite alberi di decisione

Nel seguito sono riportate le performance di diversi alberi di decisione, ognuno dei quali presenta una diversa configurazione dei parametri di learning. I parametri utilizzato sono i seguenti:

- `criterion`: Criterio con cui misurare la qualità di una suddivisione. I valori possibili sono `gini` (Gini impurity) e `entropy` (per l'information gain)
- `max_depth`: La massima profondità dell'albero. Se inizializzato a `None`, le suddivisioni sono attuate finché tutte le foglie non sono pure o contengono un numero di record minore di `min_samples_split`
- `min_samples_split`: Numero minimo di record necessario per suddividere un nodo intermedio
- `min_samples_leaf`: Numero minimo di record che le foglie devono contenere
- `class_weight`: Peso assegnato alle classi, utilizzato al momento della scelta della suddivisione migliore. Se il valore specificato è `None`, il peso di tutte le classi è considerato uguale a 1.

Nelle rappresentazioni grafiche dei modelli, un colore scuro di un nodo dell'albero corrisponde a un valore basso di `gini`/`entropy`. Al contrario, un colore chiaro indica un valore alto.

Modello 1

L'unico vincolo di questo modello riguarda la massima profondità dell'albero. Il modello nella figura 4.1 opera una prima suddivisione dei record in base al valore dell'attributo `satisfaction_level`: quelli aventi un valore maggiore di 0.465 sono classificati come `Notleft` (questa scelta risulta abbastanza intuitiva). Se invece `satisfaction_level` assume un valore minore o uguale a 0.465, allora il modello controlla anche il valore dell'attributo `number_project`: se un impiegato ha portato a termine più di 2 progetti, allora viene classificato come `Notleft`. In caso contrario, viene attribuito alla classe `Left`.

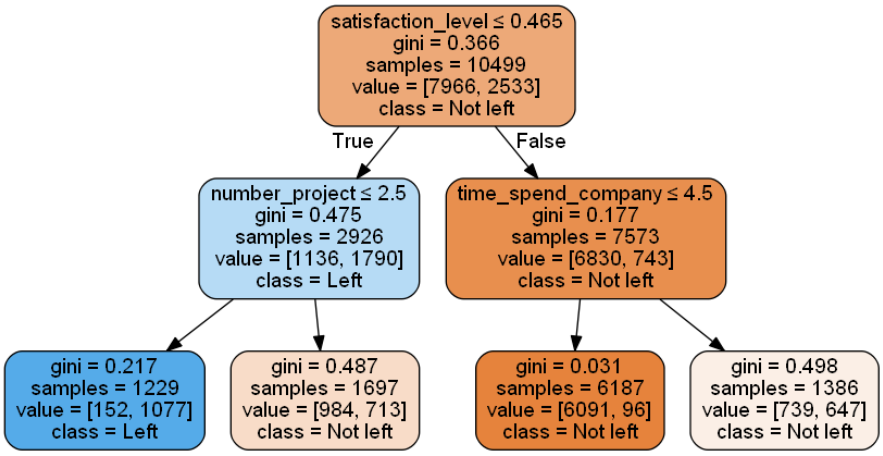


Figura 4.1: Rappresentazione grafica del modello

criterion	gini
max_depth	2
min_samples_split	2
min_samples_leaf	2
class_weight	None

Tabella 4.1: Parametri del modello

Modello 2

Per questo modello è stato deciso di rilassare leggermente il vincolo sulla profondità massima ed è stato deciso di assegnare un peso diverso a ciascuna delle due classi (in accordo alla distribuzione dei valori della feature `left`). Il modello nella figura 4.2, come il precedente, effettua inizialmente una suddivisione basandosi sull'attributo `satisfaction_level`:

- se l'impiegato ha un livello di soddisfazione inferiore o uguale a 0.465, il modello controlla `time_spend_company`. Per un valore minore o uguale a 4.5, il modello tende ad assegnare l'etichetta `Left` (la successiva divisione si basa di nuovo su `time_spend_company`). Altrimenti assegna nella maggior parte dei casi l'etichetta `Notleft` (con la successiva suddivisione su `satisfaction_level`
- se invece il livello di soddisfazione è superiore a 0.465, il modello controlla di nuovo `time_spend_company`: al contrario di prima, per un valore minore o uguale a 4.5, il modello assegna alla maggior parte dei record la classe `Notleft` (divisione successiva: `number_projects`). Per un valore maggiore di 4.5, invece, la classe è `Left` (divisione successiva: `last_evaluation`)

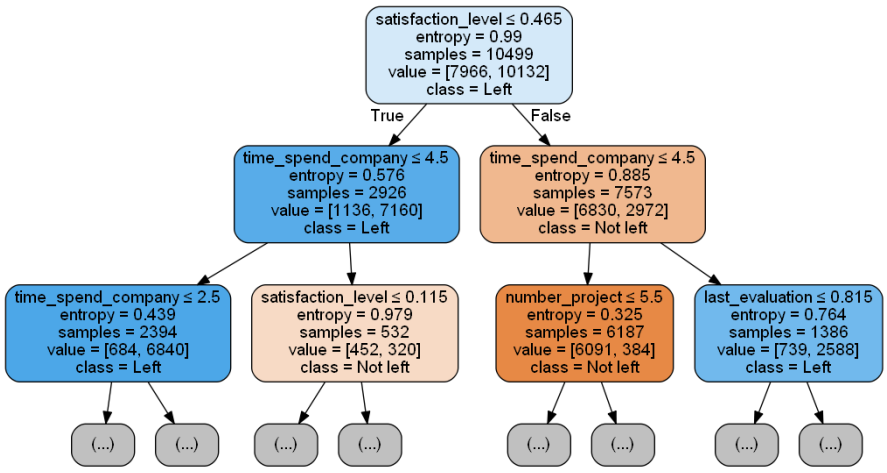


Figura 4.2: Rappresentazione grafica del modello

criterion	entropy
max_depth	3
min_samples_split	10
min_samples_leaf	10
class_weight	not left=1, left=4

Tabella 4.2: Parametri del modello

Modello 3

Anche in questo modello sono stati attribuiti pesi diversi alle due classi. Tuttavia, il vincolo sulla profondità massima è stato eliminato a favore di una strategia di *pre-pruning* (alti valori di `min_samples_split` e `min_samples_leaf`). Come nei modelli precedenti modelli, la prima suddivisione del modello in figura 4.3 riguarda `satisfaction_level`:

- `satisfaction_level` $\leq$ 0.465: all'impiegato è assegnata la classe `Left`
- `satisfaction_level` $>$ 0.465: un'ulteriore suddivisione viene effettuata in base al valore di `time_spend_company`. Se l'impiegato ha un valore dell'attributo superiore a 3.5, allora viene classificato come appartenente alla classe `Left`. In caso contrario, è molto probabile che ad egli venga attribuita l'etichetta `Notleft` (divisione successiva: `average_montly_hours`)

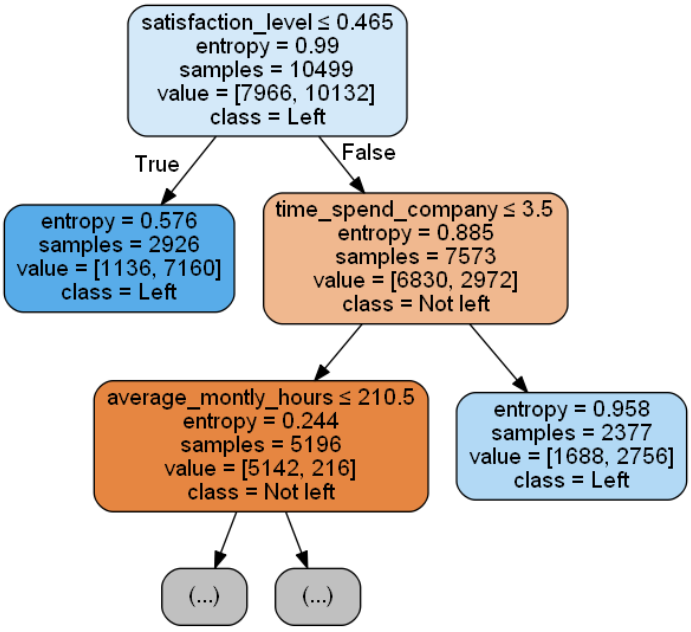


Figura 4.3: Rappresentazione grafica del modello

criterion	entropy
max_depth	None
min_samples_split	200
min_samples_leaf	2000
class_weight	not left=1, left=4

Tabella 4.3: Parametri del modello

Modello 4

L'unico vincolo di questo modello, seppur non molto stringente, riguarda `min_samples_split` e `min_samples_leaf`. Con un valore di 10 per entrambi i parametri, infatti, questa limitazione è apprezzabile soltanto ai livelli più bassi dell'albero. Nel modello in figura 4.4, ancora una volta, la prima suddivisione riguarda `satisfaction_level`:

- `satisfaction_level` $\leq$ 0.465: se il valore di `time_spend_company` è minore o uguale a 4.5, l'etichetta assegnata con più probabilità è `Left` (con la divisione successiva di nuovo su `time_spend_company`). Se `time_spend_company` è invece maggiore di 4.5, il modello tende ad assegnare l'etichetta `Notleft` (divisione successiva: `satisfaction_level`)
- `satisfaction_level` $>$ 0.465: come prima, suddividiamo in base a `time_spend_company`. Se l'impiegato ha un valore dell'attributo superiore a 4.5, allora tende ad essere classificato come appartenente alla classe `Left` (divisione successiva: `last_evaluation`). In caso contrario, è molto probabile che ad egli venga attribuita l'etichetta `Notleft` (divisione successiva: `average_montly_hours`)

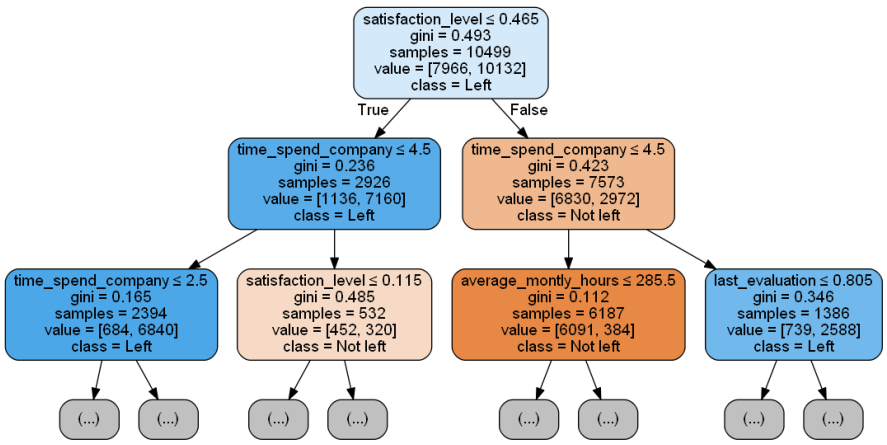


Figura 4.4: Rappresentazione grafica del modello

criterion	gini
max_depth	None
min_samples_split	10
min_samples_leaf	10
class_weight	not left=1, left=4

Tabella 4.4: Parametri del modello

4.2 Validazione dei modelli

In questa sezione si illustrano le performance dei modelli presentati nella sezione precedente. A questo scopo, il dataset originale è stato partizionato in *Traning set* (70%) e *Test set* (30%), utilizzando la tecnica del campionamento stratificato. Ogni modello è stato testato sui medesimi dati. Nella tabella 4.5 sono mostrati i risultati:

La figura 4.5 mostra le diverse matrici di confusione dei modelli. Ognuna di esse mostra il numero dei *True negative* (in alto a sx), dei *False positive* (in alto a dx), dei *False negative* (in basso a sx) e infine dei *True positive* (in basso a dx).

	Training set				Test set			
	precision	recall	f1	accuracy	precision	recall	f1	accuracy
Modello 1	0.851	0.847	0.826	0.847	0.861	0.857	0.838	0.857
Modello 2	0.926	0.914	0.917	0.914	0.922	0.908	0.911	0.908
Modello 3	0.864	0.726	0.745	0.726	0.868	0.723	0.745	0.723
Modello 4	0.973	0.971	0.972	0.971	0.962	0.960	0.961	0.960

Tabella 4.5: Performance dei modelli

4.3 Identificazione del miglior modello

Per la ricerca del miglior modello M (dal punto di vista delle performance) è stato utilizzata strategia basata su *Random Forest*. Questa soluzione consiste nel creare diversi alberi di decisione (ognuno dei quali è allenato su diversi sottoinsiemi del dataset usato per la fase di *training*). Le decisioni prese da questo modello dipendono dalle decisioni prese dai singoli alberi: in questo caso, la decisione presa dalla *Random Forest* è la decisione presa dalla maggior parte degli alberi della foresta (moda).

La scelta dei migliori parametri da utilizzare per il modello M è stata approssimata utilizzando una tecnica di *Grid search* con ricerca casuale. Tra tutte le possibili combinazioni di parametri illustrate nella tabella 4.6 (fornite tramite specifici intervalli di valori) sono state scelte casualmente 100 configurazioni e, tra di esse, è stata selezionata la migliore (vedere ancora la tabella 4.6). La funzione scelta per la misurazione della qualità delle suddivisioni è l'*Information Gain* (basata sull'`entropy`). Per la massima profondità dell'albero, invece, è stato scelto il valore 9: intuitivamente, un valore più alto avrebbe prodotto un modello con problemi di *Overfitting*. Sia per `min_samples_split` che per `min_samples_leaf` sono stati scelti due valori molto bassi: in altre parole, si è ritenuto più vantaggioso non usare tecniche di *Pre-pruning*. Parlando invece dei pesi assegnati alle classi `not_left` e `left`, notiamo che l'opzione ritenuta più efficiente è assegnare ad entrambe le classi lo stesso peso (valore `None`).

Infine, la tabella 4.7 illustra le performance di M su *Training set* e *Test set*. L'albero di decisione corrispondente è riportato nella figura 4.7: analogamente agli altri modelli studiati nella precedente sezione, vediamo che la prima suddivisione dei record avviene sull'attributo `satisfaction_level`. Al primo livello dell'albero, le suddivisioni si basano unicamente sull'attributo `number_project`: ad eccezione del modello 1 (figura 4.1), nessuno degli altri modelli aveva utilizzato questa feature nelle decisioni del primo livello. Riguardo a `Work_accident`, invece, il modello M è l'unico ad averlo coinvolto nelle decisioni dei primi due livelli.

La figura 4.6 mostra invece l'importanza di ogni attributo nel processo di decisione di M: in accordo con la prima suddivisione nella rappresentazione grafica dell'albero, vediamo che `satisfaction_level` è la feature che più incide nel processo di decisione. Seguono poi `time_spend_company` e `number_project`, mentre è evidente la poca influenza delle feature `salary` e `Work_accident`.

Allo scopo di attuare un confronto generale tra tutti i modelli precedentemente illustrati, la figura 4.8 mostra le curve ROC per tutti i modelli analizzati in precedenza.

Parametro	Intervallo	Valore scelto
criterion	['gini', 'entropy']	entropy
max_depth	[None]	9
min_samples_split	[2, 3, ..., 51]	8
min_samples_leaf	[2, 3, ..., 51]	2
class_weight	[(not_left=1, left=4), None, Balanced]	None

Tabella 4.6: Griglia dei parametri del modello M e valori selezionati

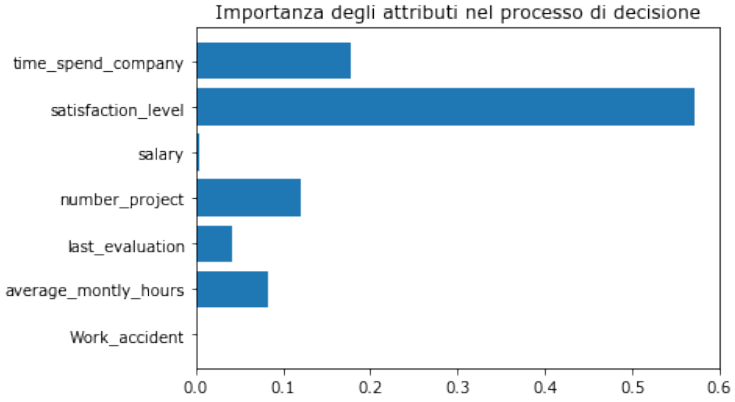
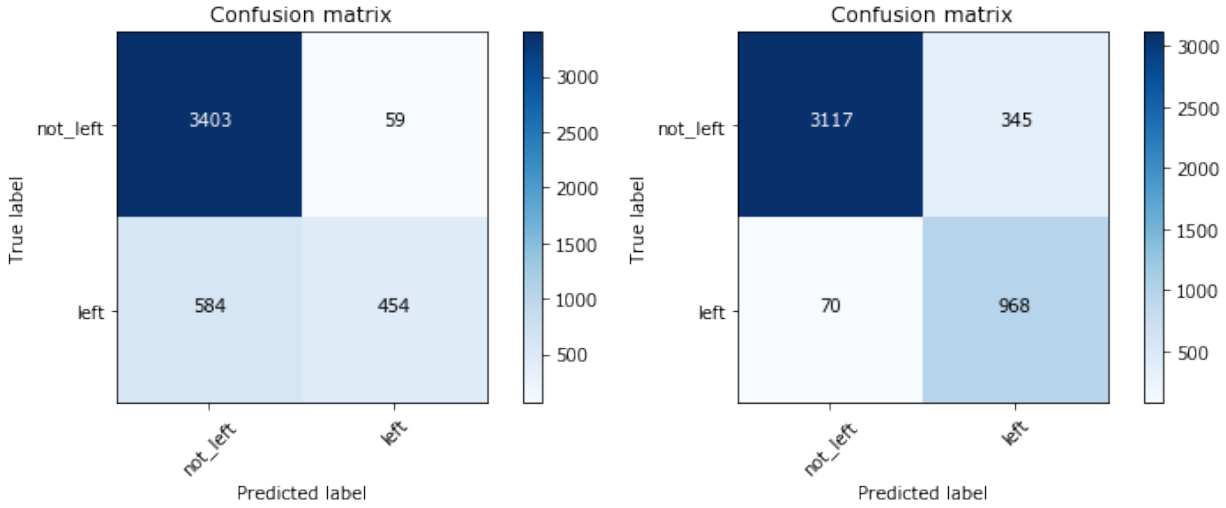
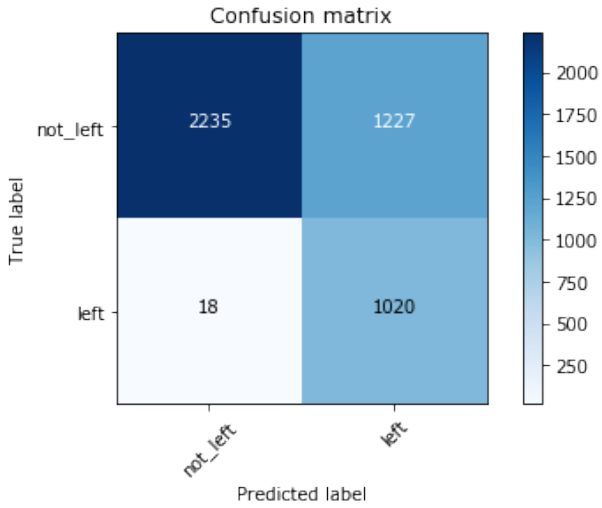


Figura 4.6: Feature importance del modello M

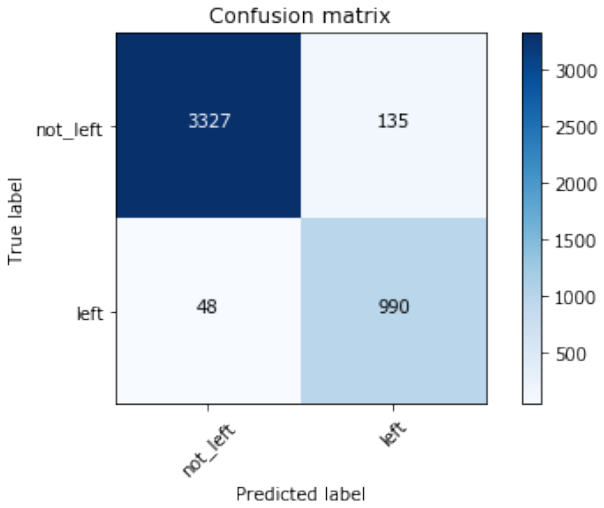


(a) Matrice di confusione del modello 1

(b) Matrice di confusione del modello 2



(c) Matrice di confusione del modello 3



(d) Matrice di confusione del modello 4

Figura 4.5: Matrici di confusione dei modelli

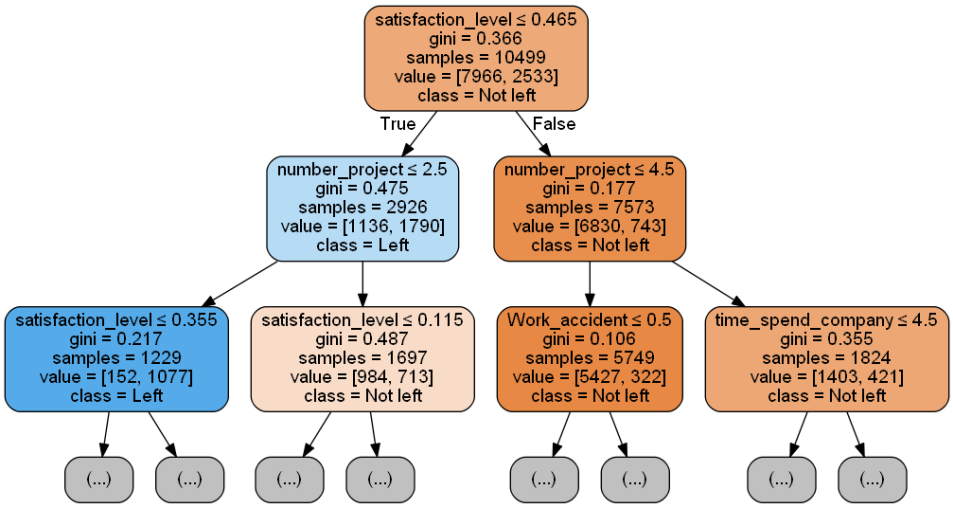


Figura 4.7: Rappresentazione grafica del modello M

Training set				Test set			
precision	recall	f1	accuracy	precision	recall	f1	accuracy
0.983	0.983	0.983	0.983	0.974	0.974	0.974	0.974

Tabella 4.7: Performance del modello *M*

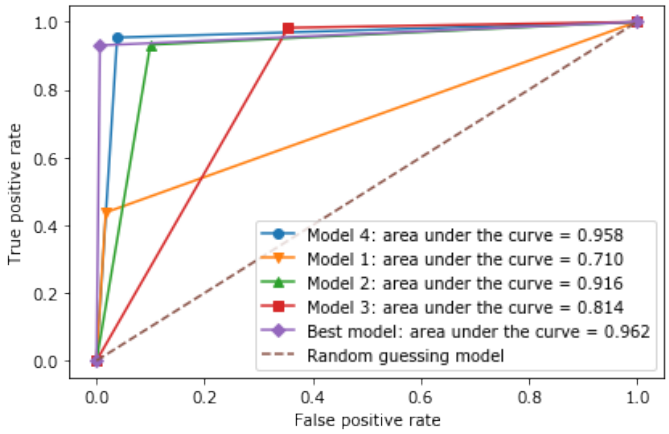


Figura 4.8: Curve ROC dei modelli precedentemente analizzati