Progetto Data Mining 1

Filippo Menis (Mat. 570669), Riccardo Affolter(Mat. 555563)

30 giugno 2018

Indice

1 Task: Data Understanding

Ognuno dei 14999 elementi appartenenti al dataset studiato rappresenta un impiegato all'interno dell'azienda. Tutti i dipendenti vengono descritti attraverso l'utilizzo di 10 attributi, riportati in seguito suddivisi per tipo:

- **Attributi binari**: Work_accident, promotion_last_5years, left
- **Attributi discreti** (numerici): number_project, time_spend_company;
- **Attributi continui**: satisfaction_level, last_evaluation, average_montly_hours;
- **Attributi categorici**: sales, salary.

I vari attributi rappresentano diversi aspetti del dipendente:

- *left*: l'attributo rappresenta il fatto che l'impiegato abbia lasciato la compagnia (1) oppure stia tutt'ora lavorando al suo interno;
- *Work_accident*: rappresenta il fatto che l'impiegato abbia avuto incidenti sul lavoro (1) o no (0);
- *promotion_last_5years*: rappresenta se l'impiegato è stato promosso negli ultimi 5 anni (1) o meno(0);
- *number_project*: rappresenta il numero di progetti a cui l'impiegato ha lavorato;
- *time_spend_company*: rappresenta il numero di anni durante i quali l'impiegato ha lavorato presso l'azienda;
- *satisfaction_level*: indica quanto un impiegato sia soddisfatto del suo lavoro;
- *last_evaluation*: indica quanto i datori di lavoro siano soddisfatti dell'impiegato;
- *average_montly_hours*: rappresenta la media di ore trascorse a lavoro durante un mese;
- *sales*: rappresenta il dipartimento a cui appartiene il lavoratore;
- *salary*: indica il salario dell'impiegato (basso, medio o alto).

1.1 Outliers and missing values

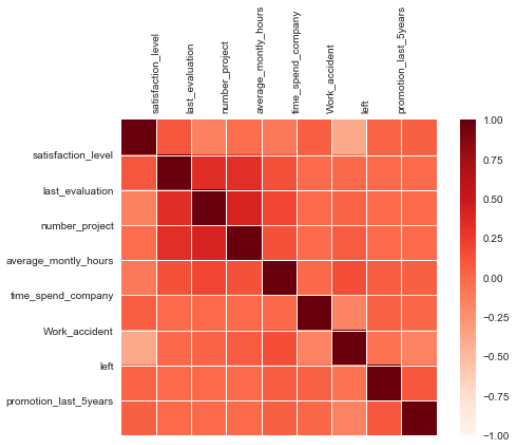
Per trovare gli outliers sono stati generati i boxplot di ogni attributo.

I risultati hanno evidenziato il fatto che tutti gli attributi tranne *time_spend_company* non evidenziano presenza di outliers. Per quanto riguarda questo attributo particolare però, è possibile affermare che la presenza di outliers non è dovuta a errori o a problemi di qualità del dataset, dato che rappresentano una piccola parte di quest'ultimo (e rappresentano i membri senior dell'azienda).

Durante una prima fase di analisi inoltre è stato possibile accertarsi del fatto che non sono presenti valori mancanti.

1.2 Pairwise Correlation

Figura 1: Correlation matrix



Dopo aver generato la correlation matrix (1) non sono stati individuate coppie di attributi la quale correlazione assumesse un valore maggiore della default threshold (pari a |.80|).

Di conseguenza, non ci sono correlazioni forti tra nessuna coppia di attributi e quindi nessun valore sarà scartato. Il valore più alto individuato è pari a 0.417211 fra gli attributi *average_montly_hours* e *number_project*.

1.3 Attributes Distribution and Statistics

In seguito viene riportata un'analisi statistica del dataset:

- *project_number*: come si può vedere nel grafico 2, il maggior numero di lavoratori ha partecipato a 3 o 4 progetti. Si può inoltre notare una distribuzione simile a quella Gaussiana.

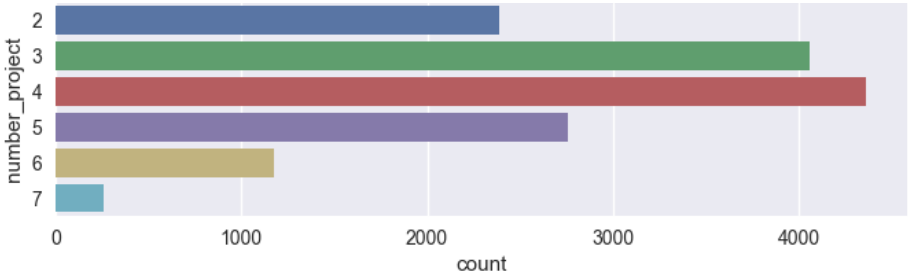


Figura 2: *number_project* distribution

- *salary*: come evidenziato nella figura 3, la maggior parte dei lavoratori riceve uno stipendio di livello medio-basso.

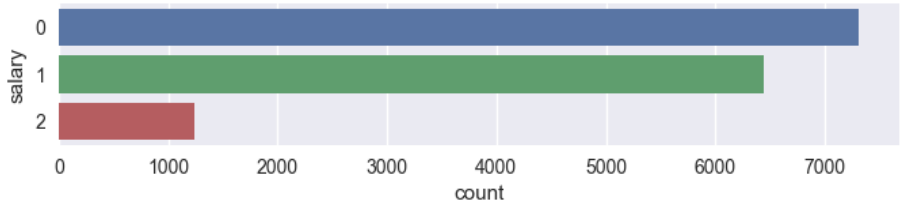


Figura 3: *salary* distribution

- *promotion_last_5years*: come evidenziato nel grafico 4, la grande maggioranza delle persone negli ultimi 5 anni non ha ricevuto una promozione.

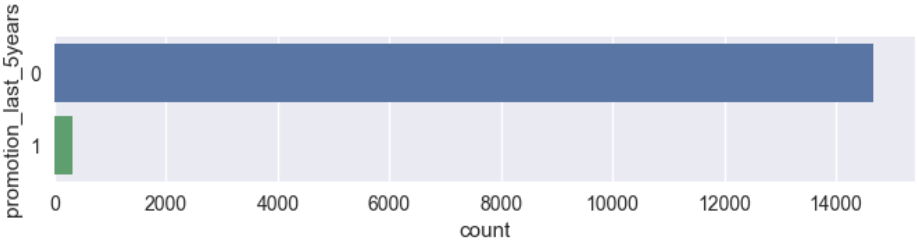


Figura 4: *promotion_last_5years* distribution

- *left*: come evidenziato nella figura 5, la grande maggioranza delle persone negli ultimi 5 anni non ha ricevuto una promozione.

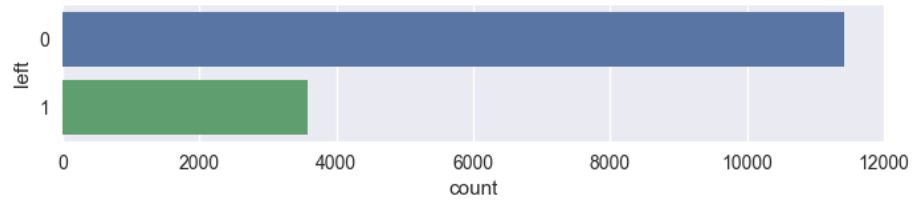


Figura 5: *left* distribution

- *sales*: come è possibile notare dal bar plot 6, le divisioni dell'azienda alle quali appartengono più lavoratori sono: sales, technical, support.

Figura 6: *sales* distribution

- *time_spend_company*: come è possibile notare dal bar plot 7, la maggior parte dei lavoratori opera all'interno dell'azienda da 3 anni, mentre un numero pari a quasi la metà di questi ultimi ci lavora da 2 anni.

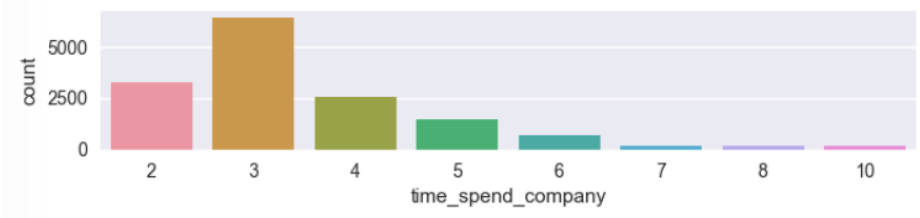


Figura 7: *sales* distribution

- *Work_accident*: dal bar plot 8 si può notare che una piccola porzione di dipendenti (poco più di 2000) ha avuto incidenti sul posto di lavoro.

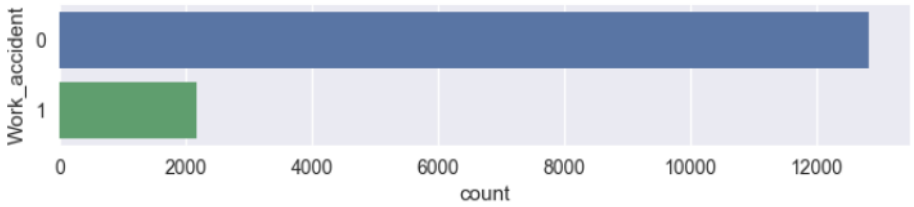


Figura 8: *Work_accident* distribution

- *last_evaluation*: il grafico 9 mostra che l'attributo segue una distribuzione bimodale. Si possono notare i due picchi intorno ai valori 0.5 e 0.9.

Figura 9: *last_evaluation* distribution

- *average_montly_hours*: dal grafico 10 si può notare come l'attributo segua una distribuzione bimodale, con i due picchi attorno ai valori 150 e 250, e la grande maggioranza dei lavoratori ricada nell'intervallo fra questi due valori.

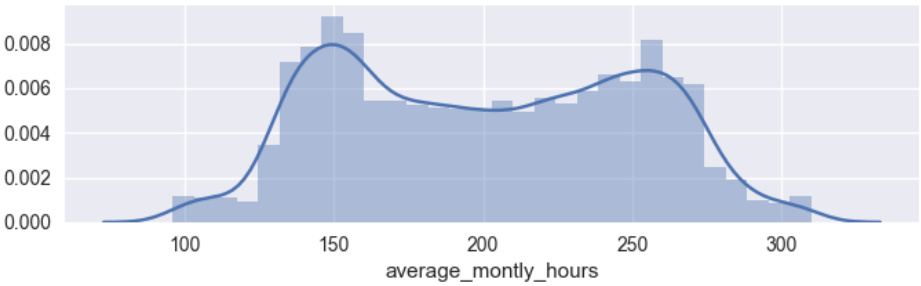


Figura 10: *average_montly_hours* distribution

- *satisfaction_level_distribution*: come possibile notare dal grafico 11, la maggior parte dei valori ricade nell'intervallo 0.4 e 1, con un eccezione per quanto riguarda i valori attorno 0.15.

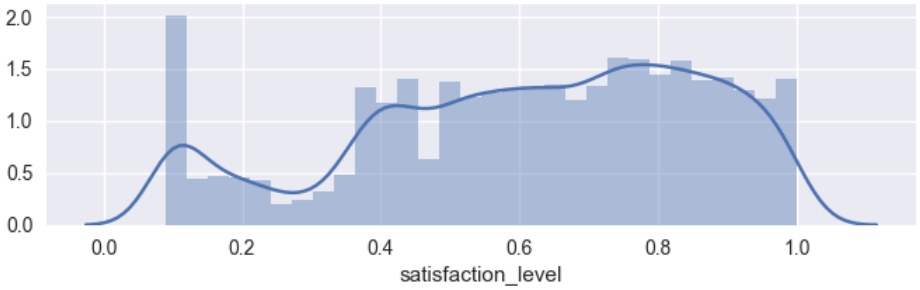


Figura 11: *satisfaction_level_distribution* distribution

Dato che l'obiettivo della ricerca sarà individuare i lavoratori che hanno lasciato l'azienda, viene riportato qui lo scatter plot rappresentante quest'ultimi (12).

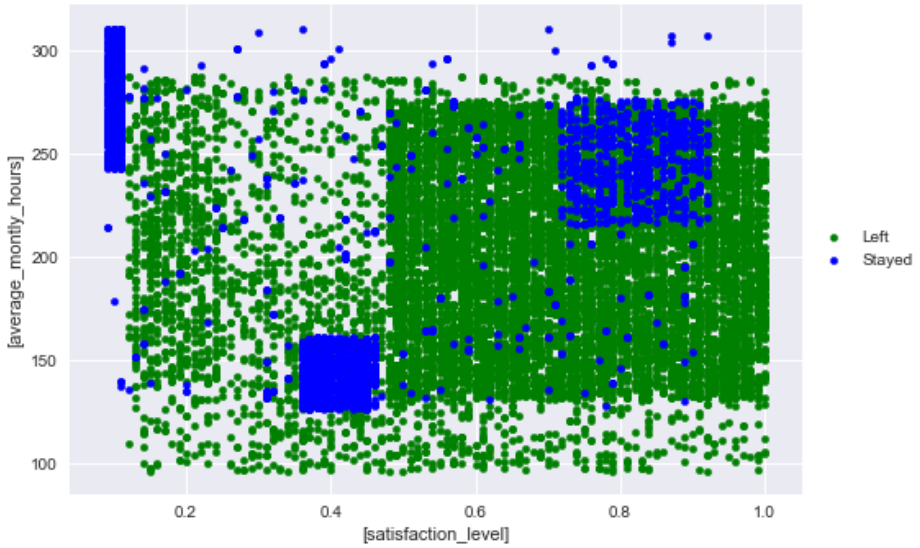


Figura 12: scatter plot

2 Task: clustering

2.1 Data preparation

Il dataset è stato modificato per permettere l'esecuzione dei vari algoritmi. L'attributo *left* è stato rimosso poiché l'esecuzione del clustering ha come obiettivo scoprire quali lavoratori hanno lasciato l'azienda.

Gli attributi *salary*,*sales*,*promotion_last_5years* e *Work_accident* sono stati rimossi poichè sono attributi categorici e binari, quindi di scarsa importanza in questo caso.

A tutti i restanti attributi è stata applicata una normalizzazione min-max così da assegnare a ognuno di essi valori compresi tra 0 e 1.

2.2 K-Means

Per trovare il valore corretto di k, è stato generato un SSE plot (13). Grazie a questo e all'utilizzo del metodo del gomito è stato possibile identificare il miglior numero di cluster da individuare, ovvero 10.

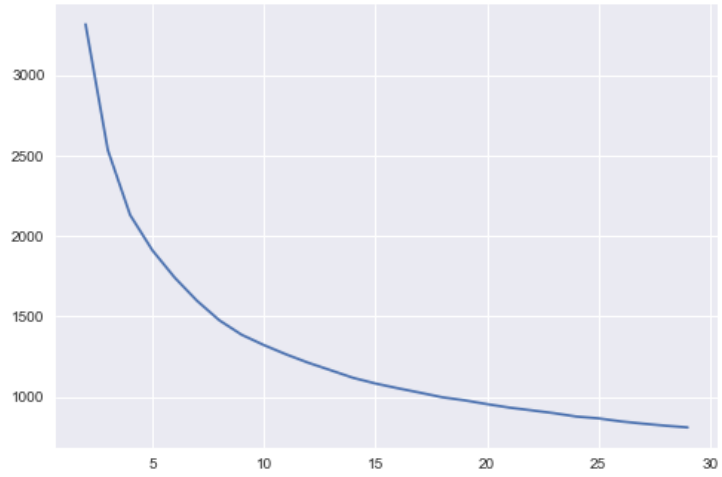


Figura 13: *SSE* plot

Il risultato dell'esecuzione di K-Means può essere osservato in figura (14). Comparando questo grafico con quello in figura (12), si può notare come l'algoritmo sia stato in grado di individuare i cluster di interesse: i gruppi color marrone (con valori di soddisfazione bassa e un alto numero di ore), lilla(con un basso numero di ore e un poca soddisfazione) e viola (alto numero di ore e soddisfazione elevata). I restanti cluster sono più sparsi e non destano interessi particolari.

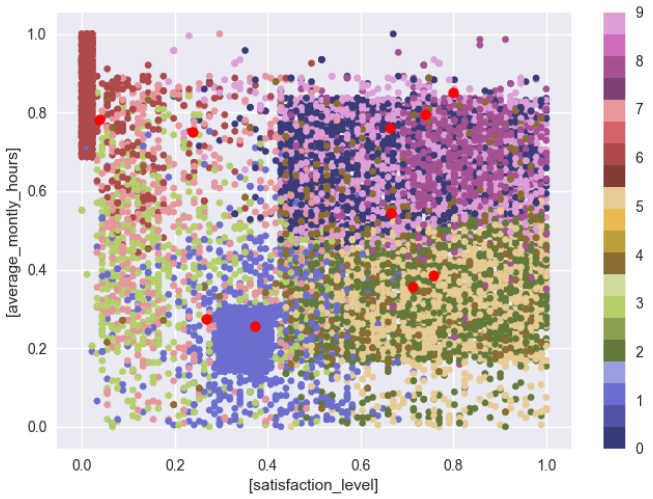


Figura 14: risultati K-Means (in rosso sono rappresentati i centroidi)

2.3 DBSCAN

Per eseguire DBSCAN è molto importante individuare i corretti valori di k ed *eps*. Il valore scelto per k è 10, poichè, in seguito a ripetute prove con diversi valori, è quello che ha ottenuto risultati migliori. Per quanto riguarda *eps* invece, il valore ottimale, pari a 0.14, è stato ottenuto osservando la curva generata dal grafico K-NN delle distanze.

Com'è possibile osservare dalla figura (15), l'algoritmo, utilizzando la distanza euclidea, non è stato in grado di individuare accuratamente tutti i cluster di interesse, ottenendo un valore di silhouette vicino a 0.

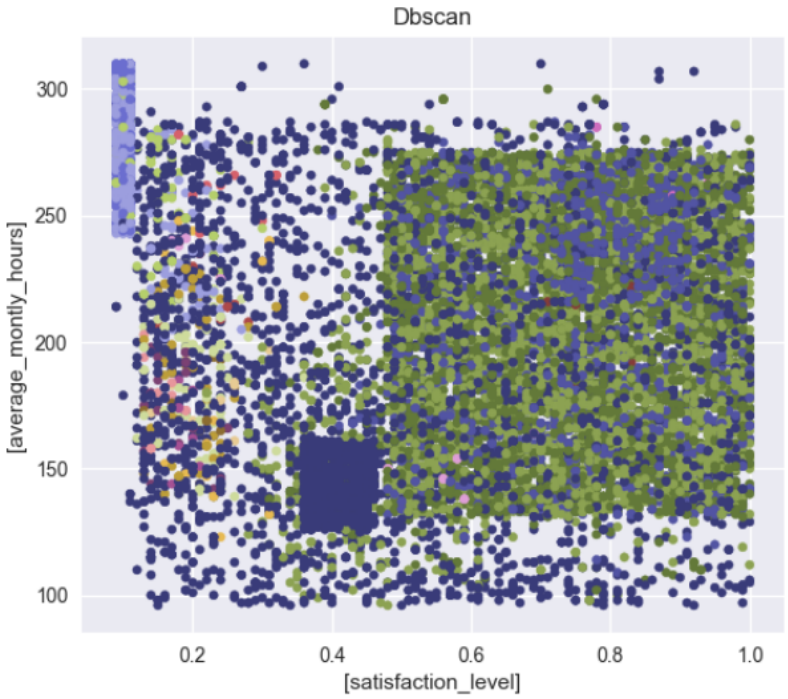


Figura 15: risultati DBSCAN

2.4 Hierarchical Clustering

Per quanto riguarda lo Hierarchical Clustering, sono state provate più combinazioni dei diversi parametri: sono state usati i metodi Ward, average, single e complete link, sia con le distanza euclidea che quella di Manhattan (tranne che per il metodo Ward dato che funziona solo con la distanza euclidea). In figura (20) sono riportati i dendrogrammi. Come si può osservare, i risultati sono piuttosto simili tranne per quanto riguarda l'approcio denominato single-link. Per motivi di chiarezza i grafici sono stati tagliati in modo da ottenere 35 cluster.

Dopo aver visualizzato tutti i risultati, è stato possibile affarmare che Ward sembra il metodo più efficace per il clustering del dataset: ha infatti un valore di silhouette più alto.

2.5 Comparison of the different clustering techniques

Come si è potuto vedere, tutti e tre gli algoritmi sono riusciti a rilevare i tre cluster di nostro interesse con risultati più o meno efficaci: mentre K-Means e lo Hierarchical Clustering hanno individuato con una buona precisione i

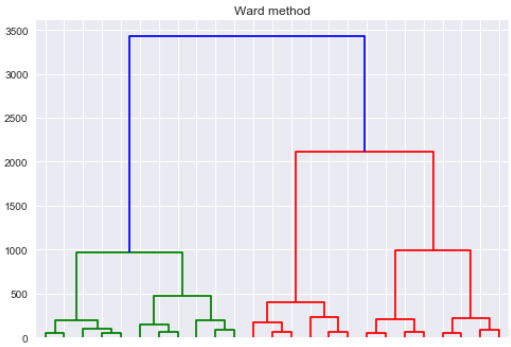


Figura 16: Ward method

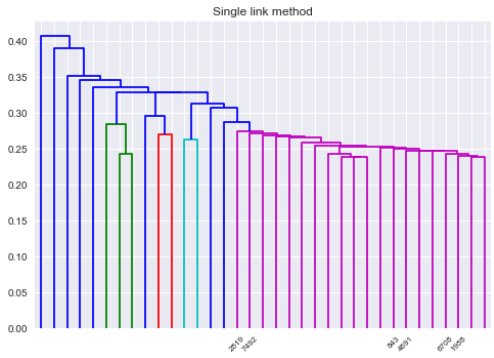


Figura 17: Single link method

Figura 18: Complete link method

Figura 19: Group average method

Figura 20: Risultati

tre cluster, DBSCAN ha ottenuto dei cluster piuttosto sparsi e di precisione minore rispetto agli altri due algoritmi. In generale, è possibile affermare che K-Means ha portato ai risultati migliori.

3 Task: Association Rules Mining

Prima di iniziare l'estrazione dei frequent pattern e delle association rules attraverso l'uso di Apriori, il dataset andava modificato. Sono state quindi apportate le seguenti modifiche alle variabili continue:

- i valori di *satisfaction_level* sono stati discretizzati in 5 intervalli, rinominati per motivi di presentazione: il risultato finale sono 5 valori

diversi denominati 'Very_low_sat', 'Low_sat', 'Medium_sat', 'High_sat' e infine 'Very_high_sat'.

- i valori di *last_evaluation* sono stati discretizzati in 5 intervalli, rinominati in seguito: 'Low_grade', 'Medium_grade','High_grade','Very_high_grade'.
- i valori di *average_montly_hours* sono stati discretizzati in 10 intervall. In seguito alla modifica i valori sono i seguenti: '[80,118]_H', '[119,140]_H', '[140,161]_H', '[162,182]_H', '[183,225]_H', '[226,247]_H', '[248,269]_H', '[270,289]_H', '[290,311]_H'

Per questo compito sono stati considerati tutti gli attributi. Oltre alle modifiche apportate ai tre a valori continui, sono state aggiunte a tutti gli altri attributi un abbreviazione simbolica per facilitare l'interpretazione dei risultati finali, com'è possibile vedere nella figura (21).

sfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	left	promotion_last_5years	sales	salary
Low_sat	Medium_grade	2_NP	[140,161]_H	3_T	0_ACC	1_L	0_PR	sales_SALES	0_SALARY
High_sat	High_grade	5_NP	[248,269]_H	6_T	0_ACC	1_L	0_PR	sales_SALES	1_SALARY
Very_low_sat	High_grade	7_NP	[270,289]_H	4_T	0_ACC	1_L	0_PR	sales_SALES	1_SALARY
High_sat	High_grade	5_NP	[183,225]_H	5_T	0_ACC	1_L	0_PR	sales_SALES	0_SALARY
Low_sat	Medium_grade	2_NP	[140,161]_H	3_T	0_ACC	1_L	0_PR	sales_SALES	0_SALARY

Figura 21: Il dataset ottenuto

3.1 Frequent Patterns

Per estrarre i frequent pattern è stato utilizzato Apriori. Sono stati analizzati solo pattern con una lunghezza minima di 3. Sono stati valutati 3 valori di support: 10%, 20% e 30% del numero totale di elementi appartenenti al dataset. Il numero di pattern trovati è riportato nella tabella (1).

Support	Frequent I.	Closed I.	Max I.
10%	213	213	79
20%	38	38	19
30%	9	9	9

Tabella 1: Numero di pattern di lunghezza 3 estratti con diversi valori di support

Inoltre nella tabella (2) vengono riportati i frequent patterns e i maximal frequent itemset (i closed non sono stati considerati poiché non mostrano caratteristiche interessanti).

Support	Frequent I.	Max I.
10%	1. Medium_sat, 0_L, 0_ACC, 0_PR 2. Very_high_sat, 0_L, 0_ACC 3. High_sat, 0_L, 0_PR, 2790 4. Very_high_sat, 0_L, 0_ACC, 0_PR 5. High_grade, 0_SALARY, 0_PR	1. Medium_sat, 0_L, 0_ACC, 0_PR 2. Very_high_sat, 0_L, 0_ACC, 0_PR 3. Medium_grade, 0_L, 0_ACC, 0_PR 4. 2_T, 0_L, 0_ACC, 0_PR 5. [183, 225]_H, 0_L, 0_ACC, 0_PR
20%	1. 1_SALARY, 0_L, 0_ACC 2. 0_SALARY, 0_L, 0_ACC 3. 0_SALARY, 0_L, 0_ACC, 0_PR 4. 1_SALARY, 0_L, 0_ACC, 0_PR 5. High_grade, 0_L, 0_ACC	1. 0_SALARY, 0_L, 0_ACC, 0_PR' 2. 1_SALARY, 0_L, 0_ACC, 0_PR 3. High_grade, 0_L, 0_ACC, 0_PR 4. 3_T, 0_L, 0_ACC, 0_PR 5. Medium_sat, 0_ACC, 0_PR
30%	1. 0_L, 0_ACC, 0_PR 2. 0_SALARY, 0_ACC, 0_PR 3. 3_T, 0_ACC, 0_PR 4. 1_SALARY, 0_ACC, 0_PR 5. 0_SALARY, 0_L, 0_PR	1. 0_L, 0_ACC, 0_PR 2. 0_SALARY, 0_ACC, 0_PR 3. 3_T, 0_ACC, 0_PR 4. 1_SALARY, 0_ACC, 0_PR 5. 0_SALARY, 0_L, 0_PR

Tabella 2: Risultati

Dato che il dataset è fortemente sbilanciato a favore dei dipendenti che sono rimasti nell'azienda (il 76% del totale), per trovare quelli che hanno lasciato l'organizzazione bisogna osservare i pattern ottenuti con un valore di support pari a 10.

In generale, osservando i primi 5 campioni di ogni insieme mostrati in tabella (2) si può affermare che tendenzialmente le persone che non hanno avuto incidenti, hanno un salario medio basso, e hanno un alto valore di soddisfazione non hanno lasciato l'azienda.

3.2 Association Rules

Per trovare le association rules sono state provate diverse combinazioni di valori di support (10%, 20%) e diversi valori di confidence (60%, 70%, 80%). Dato che mantenendo il valore di support pari al 20% non si ottiene nessuna regola contenente lavoratori che hanno lasciato l'azienda (quindi l'obiettivo che si sta cercando), saranno analizzati solo i casi nei quali $supp = 20\%$.

Sono state inoltre considerate solo le regole aventi un valore di $lift > 1.05$ poiché questo assicura di ottenere solo le regole più importanti. I risultati principali che coinvolgono esclusivamente le regole che presentano i valori $Left = 1/0$ nella parte destra, sono riportati in tabella (3).

Confidence	Left = 0	Left = 1	Lift > 1.05 & Left = 0	Lift > 1.05 & Left = 1
60%	188	8	95	8
70%	153	4	95	4
80%	95	4	95	4

Tabella 3: Risultati con $supp = 20\%$

Come si può osservare, un piccolo numero di regole, dal valore di confidence piuttosto alto, implica il fatto che un dipendente lasci l'azienda.

Regola	Confidence	Lift
{2_NP, 3_T, 0_ACC, 0_PR} $->$ {1_L}	84.31%	3.54
{2_NP, 3_T, 0_ACC} $->$ {1_L}	84.25%	3.53
{2_N, 3_T, 0_PR} $->$ {1_L}	82.43%	3.46
{2_NP, 3_T}$->$ {1_L}	82.41%	3.46

Tabella 4: Regole con $Left = 1$

Per quanto riguarda le regole contenenti nella parte destra il valore {1_L}, è interessante notare il fatto che hanno tutte un valore di $Lift$ piuttosto alto: le regole in tabella (4) presentano infatti valori attorno al 3.50, e anche le restanti quattro regole (non riportate nella tabella), ottenute con un valore di confidence minima pari a 60, hanno un valore di $Lift$ maggiore di 2.5.

Ad esempio, la regola {2_NP, 3_T} $->$ {1_L} implica che se un lavoratore ha lavorato su due progetti ed è da 3 anni nell'azienda, lascerà l'azienda. Bisogna però sottolineare che, poiché queste regole sono state estratte da un dataset sbilanciato, non sono efficacemente applicabili per costruire un classificatore basato su queste regole: su 1854 impiegati a cui questa regola è applicabile, in 326 casi si verificherà un falso positivo. Questo comporta un'accuratezza pari all' 82%. Allo stesso tempo però questa regola non è in grado di riconoscere i restanti 1.717, il che porta a bassi valori di recall e accuratezza.

Regola	Confidence	Lift
{Very_high_sat, 3_T}− > {0_L}	99.8%	1.31
{3_NP, 3_T, 0_PR} − > {0_L}	99.4%	1.30
{High_grade, 3_T, 0_PR}− > {0_L}	98.7%	1.29
{2_T, 0_SALARY, 0_PR} − > {0_L}	98.5%	1.29
{Medium_sat, High_grade} − > {0_L}	97.1%	1.27

Tabella 5: Regole con $Left = 0$

Le regole estratte contenenti {0_L} nella parte destra presentano invece una situazione lievemente diversa: come possibile osservare dalla tabella (5), molte regole hanno un altissimo valore di confidence, mentre il valore di Lift è inferiore rispetto alle regole che implicano {1_L}. In ogni caso, sono regole valide e da prendere in considerazione visto che il valore di Lift è più alto della soglia minima selezionata.

Per esempio {Very_high_sat, 3_T}− >{0_L}, la regola coi maggiori valori di confidence e lift fra quelle ottenute, la quale implica che un lavoratore molto soddisfatto e impiegato da 3 anni non lascerà l'azienda) è applicabile a 6443 impiegati, dei quali 4857 sono realmente rimasti nell'organizzazione, ottenendo quindi un valore di accuratezza del 75%.

4 Task: Classification

4.1 Feature Selection

Prima di procedere alla generazione degli alberi per la classificazione, sono stati eliminati dal dataset gli attributi considerati irrilevanti o di scarsa importanza: *sales, salary, Work_accident, promotion_last_5years* sono stati ignorati, mentre *left* è stato utilizzato come obiettivo per la classificazione.

Gli attributi utilizzati sono riportati in tabella (6) assieme al loro relativo valore di importanza per i due criteri utilizzati (Gini ed Entropy).

Attributo	Gini	Entropy
satisfaction_level	0.531	0.38
last_evaluation	0.145	0.134
number_project	0.099	0.191
average_montly_hours	0.0790	0.061
time_spend_company	0.153	0.229

Tabella 6: Attributi per la classificazione e relativa importanza (rispetto a un decision tree di profondità 6)

4.2 Decision Trees and Validation

Sono state utilizzate diverse combinazioni di metriche e profondità massima dell'albero di decisione. Quindi, utilizzando sia Gini che Entropy, sono state provate diverse profondià, pari ai valori 4, 5, 6, 7. Questo pre-pruning si è rivelato essere essenziale per evitare il fenomeno dell'overfitting.

Poiché non si aveva a disposizione un vero e proprio test set, sono stati utilizzati due approcci:

- Holdout validation: il dataset è stato suddiviso in due parti. La prima, composta dal 70% dei lavoratori, è stata usata come training set mentre il restante 30% è stato utilizzato come test set.

- Cross validation: il processo prevede la divisione del dataset in 10 parti eguali, una di queste verrà utilizzata come test set mentre le restanti 9 come training set. Questo processo è ripetuto altre volte in modo da utilizzare ogni singola parte come test set. Viene poi fatta una media dei risultati ottenuti come, ad esempio, l'accuracy.

I risultati ottenuti, comprendenti i valori di accuracy, precision, recall, f-measure per quanto riguarda il processo di holdout validation e accuracy per la cross validation, sono riportati nelle due tabelle (7), per quanto riguarda i decision tree ottenuti applicando Gini, e (8), per i decision tree generati utilizzando Entropy.

Profondità	Precision	Recall	Accuracy	F-Measure	Accuracy C.V.
4	96.9%	0.969	0.969	0.966	96.6%
5	97.5%	0.975	0.975	0.975	97.3%
6	97.8%	0.978	0.978	0.978	97.5%
7	98.1%	0.981	0.981	0.981	98.0%

Tabella 7: Risultati ottenuti con Gini con vari livelli di profondità

Profondità	Precision	Recall	Accuracy	F-Measure	Accuracy C.V.
4	96.9%	96.6%	96.6%	96.6%	96.4%
5	97.3%	97.3%	97.3%	97.3%	97.1%
6	97.7%	97.7%	97.6%	97.7%	97.5%
7	98.0%	98.0%	98.0%	98.0%	97.8%

Tabella 8: Risultati ottenuti con Entropy con vari livelli di profondità

4.3 Interpretation

In questa sezione verranno analizzati i risultati ottenuti dal decision tree (figura 22) generato utilizzando il criterio Gini e 4 come profondità massima, in modo da poterlo analizzare comodamente. Come si può vedere dalla tabella (7), pur non essendo profondo, il decision tree ha ottenuto alti valori per tutte le metriche considerate (attorno al 97%).

Come si può notare dalla figura (22), le due categorie principali di lavoratori che hanno lasciato l'azienda sono quelli che:

- presentano un basso livello di soddisfazione, hanno lavorato su pochi progetti (meno di 2) e non sono stati valutati molto positivamente (last_evaluation $<= 0.445$).
- presentano un basso livello di soddisfazione, hanno lavorato su più di 2 progetti ma sono molto insoddisfatti (satisfaction_level $<= 0.115$).
- presentano un alto livello di soddisfazione, sono stati valutati positivamente e hanno lavorato meno di 216 ore, in media, al mese.

Inoltre è interessante notare come 8821 lavoratori ricadano nella stessa foglia: 8706 di questi sono rimasti nell'azienda e presentano un alto livello di soddisfazione, meno di 4 anni all'interno della compagnia, meno di 290 ore medie al mese e hanno lavorato a meno di 6 progetti.

4.4 Comparison

Come mostrato nelle tabelle (7) e (8), le differenze fra i diversi risultati sono minime. All'aumentare della profondità dell'albero si ottengono dei miglioramenti sotto ogni punto di vista, tranne per quanto riguarda la chiarezza dell'albero e il rischio di incappare nell'overfitting.

La curva ROC (Receiver Operating Charachteristic) (figura 23) rappresenta le performance del decision tree mostrato in precedenza (figura 22). Intuitivamente, più la curva è vicina all'angolo in alto a sinistra, più alta sarà la sua AUC (Area Under Curve), migliore saranno i risultati del decision tree. In questo caso la AUC ottenuta è pari a 0.94, il che è un valore piuttosto alto (1 rappresenta un classificatore perfetto).

Le performance del decision tree con profondità 4 (figura 22) sono riportati in figura (24).

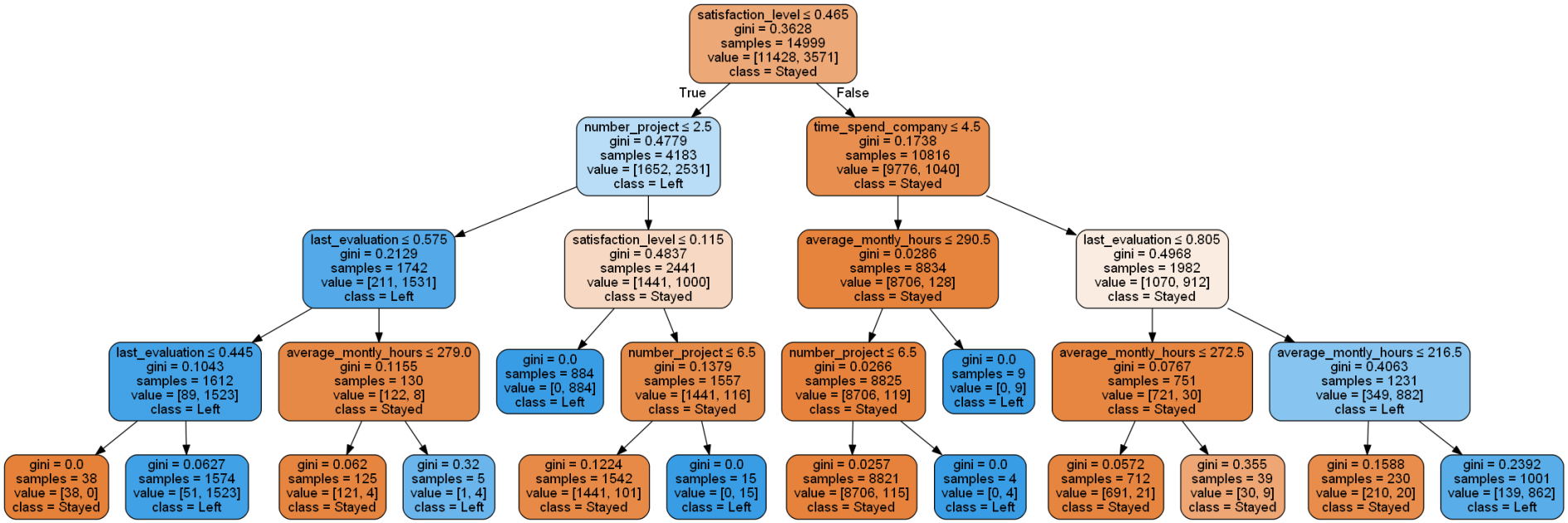


Figura 22: Decision tree, depth = 4, Gini

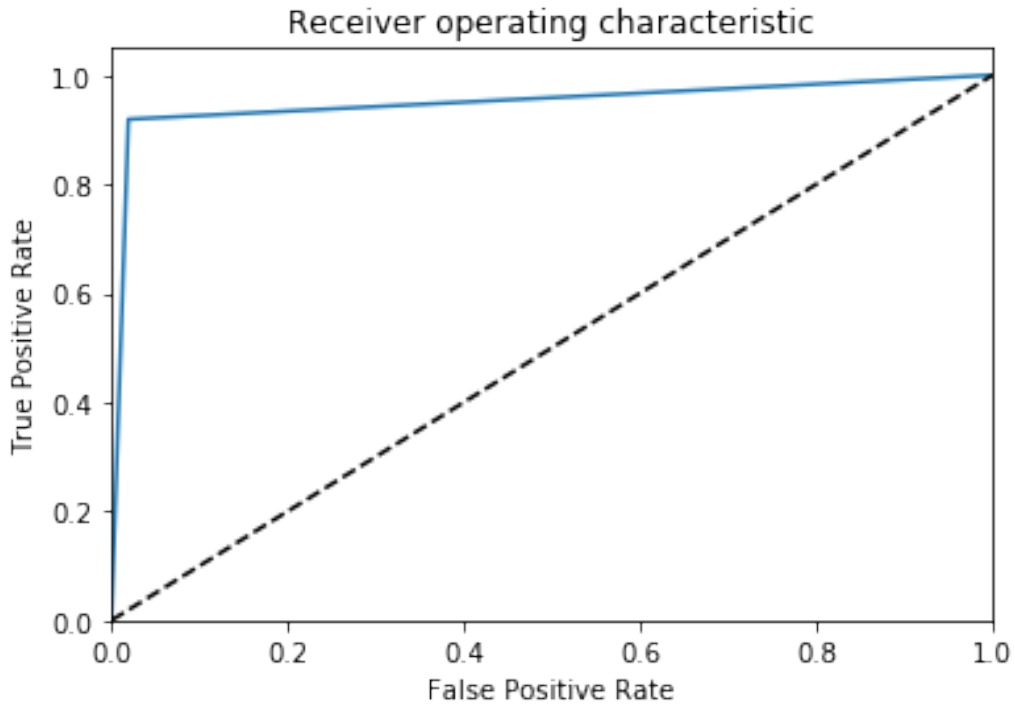


Figura 23: La curva ROC del decision tree in (22)

	precision	recall	f1-score	support
Not Left	0.98	0.98	0.98	3462
Left	0.93	0.92	0.93	1038
avg / total	0.97	0.97	0.97	4500

Figura 24: Risultati del decision tree in (22)

In generale, visto il fatto che i decision tree (con ogni combinazione di profondità e criterio di splitting) sono stati in grado di rilevare con precisione i lavoratori che hanno abbandonato l'azienda, pur essendo il dataset pesantemente sbilanciato verso quelli che sono rimasti, si può affermare che per questo dataset il miglior metodo di classificazione è un decision tree.