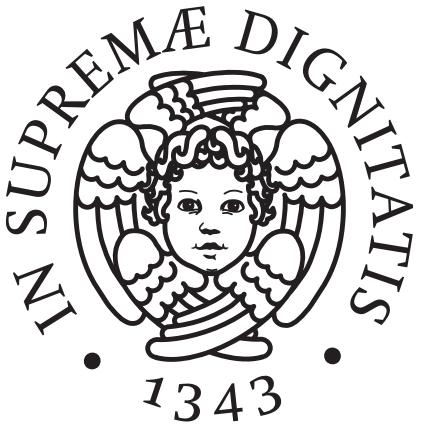


Università di Pisa

Progetto per il corso di Data Mining
A.A. 2017/2018

Analisi del Dataset
Human Resources Analytics

Gianmarco Ricciarelli
Maria Cristina Uccheddu
Stefano Carpita

Indice

1 | Data Understanding

1.1 Obiettivi

In questo progetto viene analizzato il dataset (simulato) *Human Resources Analytics* contenente le informazioni sui dipendenti di un'azienda fittizia. Come mostrato in Figura 1.1 su un totale di 14999 dipendenti il 24%, corrispondente a 3571 lavoratori, ha lasciato l'azienda. Gli obiettivi primari dell'analisi sono i seguenti:

- capire i motivi principali per cui i lavoratori hanno lasciato l'azienda;
- predire probabilisticamente se un lavoratore lascerà in futuro l'azienda;
- indicare al management dell'azienda dei provvedimenti da attuare per ridurre il numero di impiegati che la abbandonano.

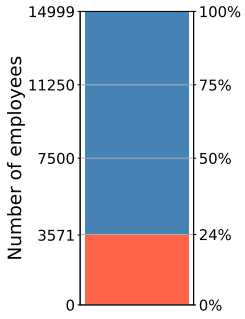


Figura 1.1: Numero di lavoratori

1.2 Data semantics

In tabella 1.1 sono riportate le variabili relative ai dipendenti dell'aziendae la corrispondente tipologia.

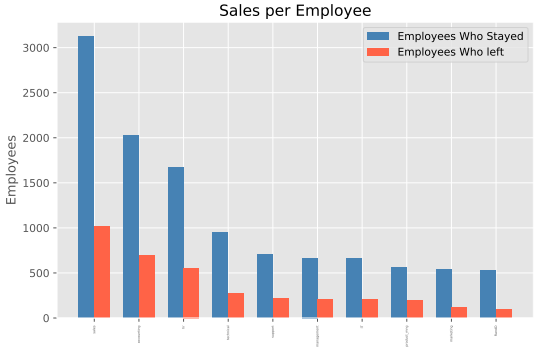
Con la prima variabile, *Satisfaction Level*, viene fornita una valutazione quantitativa del livello di soddisfazione di ciascun dipendente, in un range che va da un valore minimo di 0 ad un massimo di 1. *Last Evaluation* fornisce il tempo trascorso, in anni, dall'ultima valutazione delle performace del dipendente. *Number Project* riporta il numero di progetti completati da ciascun dipendente durante il periodo di lavoro. *Average Montly Hours* rappresenta la media delle ore di lavoro in un mese. *Time Spend Company* corrisponde al numero di anni trascorsi dal dipendente all'interno dell'azienda. *Work Accident* esprime con un 1 il coinvolgimento di un dipendente in un incidente sul lavoro, altrimenti viene impostato come 0. *Left* è utilizzata per tener traccia dei dipendenti che hanno lasciato l'azienda, per i quali viene usato il valore 1, mentre per quelli che sono rimasti viene usato il valore 0. *Promotion last 5 Year* esprime con un 1 se il dipendente è stato promosso negli ultimi 5 anni, altrimenti assume il valore 0. *Sales* definisce il dipartimento nel quale il dipendente lavora, e *Salary* esprime il livello (*low*, *medium*, *high*) nel quale rientra il salario del dipendente. Le descrizioni delle variabili sono state estrapolate dai metadati forniti assieme al Dataset sulla pagina di Kaggle [1] nella quale il Dataset è contenuto. La scelta di catalogare le colonne *Work Accident*, *Left* e *Promotion Last 5 Years* come variabili di tipo categorico verrà opportunamente motivata nelle sezioni successive.

Field	Type
satisfaction_level	continuous
last_evaluation	continuous
number_project	discrete
average_montly_hours	discrete
time_spend_company	discrete
salary	ordinal
work_accident	categorical
promotion_last_5years	categorical
sales	categorical
left	categorical

Tabella 1.1: Variabili presenti nel Dataset e rispettivi tipi.

1.3 Distribution of the variables and statistics

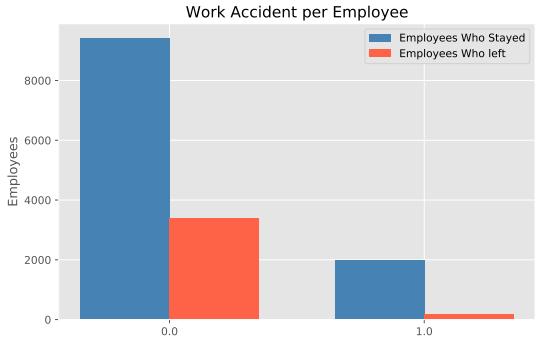
In questo paragrafo vengono presentati i grafici relativi alla distribuzione dei valori assunti dalle variabili descritte nella sezione precedente. Per dare una migliore interpretazione a questi abbiamo deciso di sostenere un'analisi accurata che contraddistingue i dipendenti che lavorano nell'azienda, rappresentati dal colore blu, e quelli che invece la hanno lasciata, rappresentati dal colore rosso. Prima di tutto vogliamo studiare la distribuzione dei dipendenti rispetto alle variabili categoriche: Divisione dipendenti

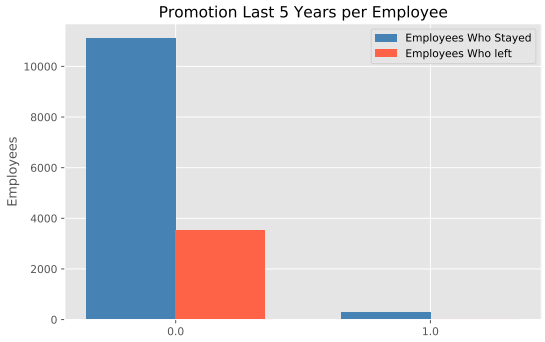


totali per left , Come si può constatare dal grafico sulla sinistra in un totale di 14999 dipendenti analizzati: 3571 di questi hanno lasciato l'azienda (quasi il 24% del totale), mentre i restanti 11428 dipendenti

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

(quasi il 76%) sono rimasti nella azienda. La distribuzione mostrata sulla destra invece indica la distribuzione degli impiegati nei vari settori: (parte mancante) .



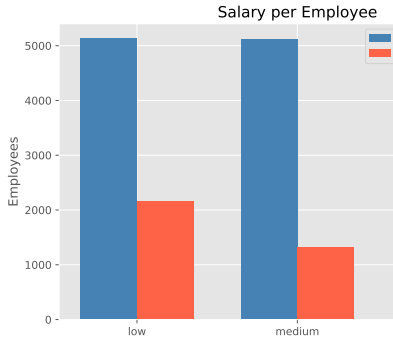


Nel grafico a sinistra si studia il rapporto tra i dipendenti e la presenza o meno di un infortunio durante il periodo di lavoro all'interno dell'azienda e si è riscontrato che soltanto 169 impiegati tra i 3571 che se ne sono andati hanno avuto un incidente sul lavoro, mentre gli impiegati ancora all'interno dell'azienda ad aver subito un incidente sono 2000 su 11428.

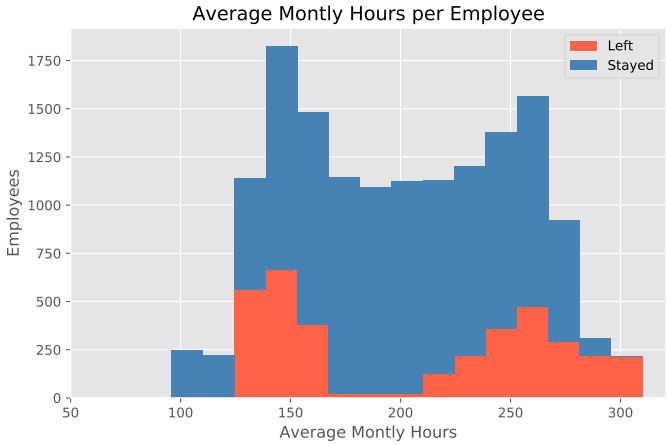
Nel grafico a destra invece rapportiamo ciascun dipendente al fatto che questo sia stato promosso negli ultimi 5 anni oppure no, possiamo ricavare una informazione importante, la totalità degli impiegati che hanno lasciato l'azienda non ha avuto una promozione negli ultimi 5 anni. Soltato 19 impiegati su 11428 di quelli rimasti sono stati promossi.

Una volta studiate le distribuzioni categoriche continuiamo l'analisi con gli altri attributi, per renderle più esplicative abbiamo affiancato a ciascun grafico la tabella che lo descrive in modo accurato e in modo che fosse autoesplicativo:

Abbiamo rapportato in primis i dipendenti con il proprio salario, le percentuali che vengono indicate sul grafico non sono in base alla totalità dei dipendenti ma riguardano solo il tipo di dipendente definito dalla riga di appartenenza.



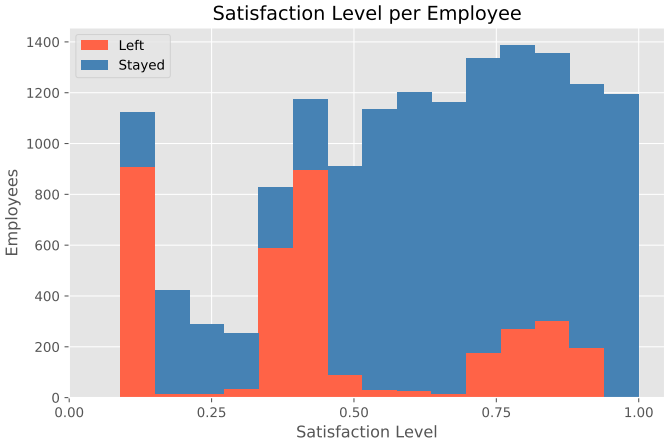
Proseguendo poi con il rapporto tra dipendenti ancora in azienda e non, e il numero delle ore di lavoro in media:



Dipendente	Low	Medium	
InAzienda	5144 (circa 45.02%)	5129 (circa il 44.88%)	11
OutAzienda	2172 (circa 60.8%)	1317 (circa il 36.9%)	

Tabella 1.2: Distribuzione salario per dipendente.

A questo punto è giusto analizzare il livello di soddi-

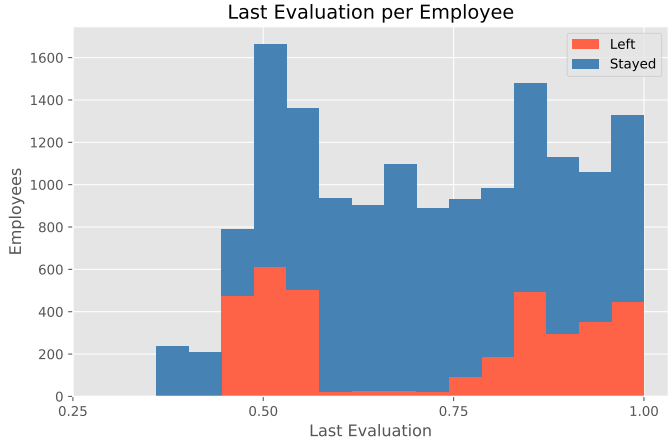


sfazione e possiamo constatare dalla tabella che:

Dipendente	media	dev. std.	min,max
InAzienda	199.06	45.68	96 , 287
OutAzienda	207.42	61.20	126 , 310

Tabella 1.3: Il tempo medio di ore di lavoro al mese per dipendente

E la possiamo rapportare rispetto alla precedente valutazione del livello di soddisfazione (cambiare valori della tabella)



Dipendente	media	dev. std.	min,max
InAzienda	0.67	0.22	0.12 , 1.0
OutAzienda	0.44	0.26	0.09 , 0.92

Tabella 1.4: Livello di soddisfazione per il dipendente

1.4 Data quality

- Missing values

- Outliers

L'individuazione dei possibili outliers di una variabile numerica consiste nel verificare se siano presenti dei valori estremi rispetto alla distribuzione dei dati osservati. I test comunemente utilizzati, come il test di Grubb o il criterio di Chauvenet, sono basati sull'assunzione di una distribuzione di probabilità gaussiana, che non si osserva per le variabili numeriche del dataset analizzato (spiegare in distribution of the variables). Un metodo robusto e di immediata applicazione è quello di osservare il boxplot dei dati, identificando come candidati outliers i valori che si trovano al di fuori dei whiskers, ovvero valori x della variabile osservata per cui $|x - \tilde{x}| > 2\,IQR(x)$, dove $\tilde{x}$ è la mediana ed $IQR(x)$ lo scarto interquartile.

1.5 Variable transformations

Analizzando il significato delle variabili presenti nel dataset, abbiamo deciso di rappresentare *Work Accident* e *Left* utilizzando il tipo categorico piuttosto che quello discreto. Questa scelta è stata motivata dall'analisi semantica delle due variabili, le quali forniscono una

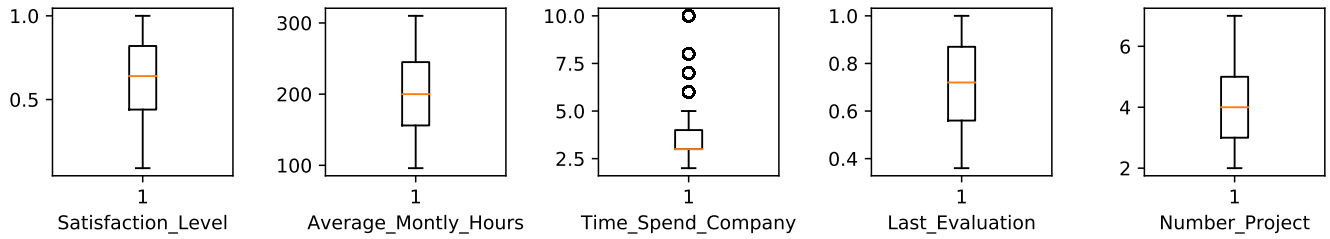
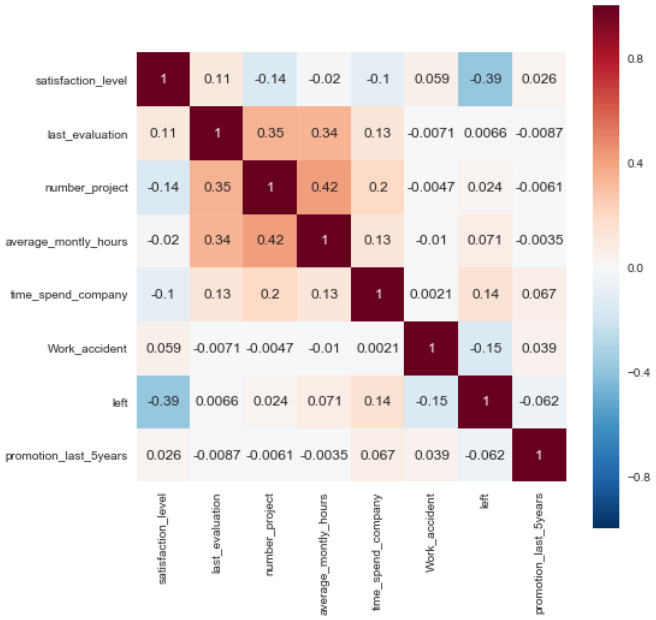


Figura 1.2: Boxplots per le variabili numeriche

risposta del tipo ”Sì o No” alle domande relative agli incidenti sul lavoro e all’abbandono o meno dell’azienda da parte dei dipendenti.

1.6 Pairwise correlations and eventual elimination of redundant variable

In questa sezione abbiamo studiato la correlazione ovvero la relazione lineare tra i vari attributi continuous o discreti.



Dalla matrice possiamo rilevare se ci sia una correlazione positiva, nulla o negativa. Sia per quanto riguarda la correlazione positiva sia per quella negativa si caratterizzano in settori: con valori da 0 a 0.3 correlato debolmente, da 0.3 a 0.7 moderatamente o maggiore di 0.7 fortemente (rispettivamente per la negativa i segni saranno negativi). Da questo possiamo definire che ad avere una correlazione debole è la variabile time_spend_company con left, last_evaluation, number_project e average_monthly_hours. Queste ultime, ad eccezione di left, invece sono correlate fra loro in modo moderato con un valore massimo di 0.42 tra average_monthly_hours e number_project. Il valore 1 indica la correlazione con se stesso che infatti è massima. Dal punto di vista della correlazione negativa, abbiamo debolmente correlati left con work_accident e satisfaction_level con number_projecy e time_spend_company. Abbiamo invece una correlazione negativa moderata tra left e satisfaction_level di valore -0.39.

2 | Clustering analysis

2.1 Clustering Analysis by K-means

2.1.1 Choice of attributes and distance function

Le variabili sulle quali abbiamo deciso di applicare la Cluster Analysis tramite K-means sono le due variabili di tipo continous presenti nel Dataset, ossia *Satisfaction Level* e *Latest Evaluation*. Le variabili di tipo categorico sono state scartate al momento della scelta dato che la natura stessa dell'algoritmo prevede il suo utilizzo su variabili di tipo numerico.

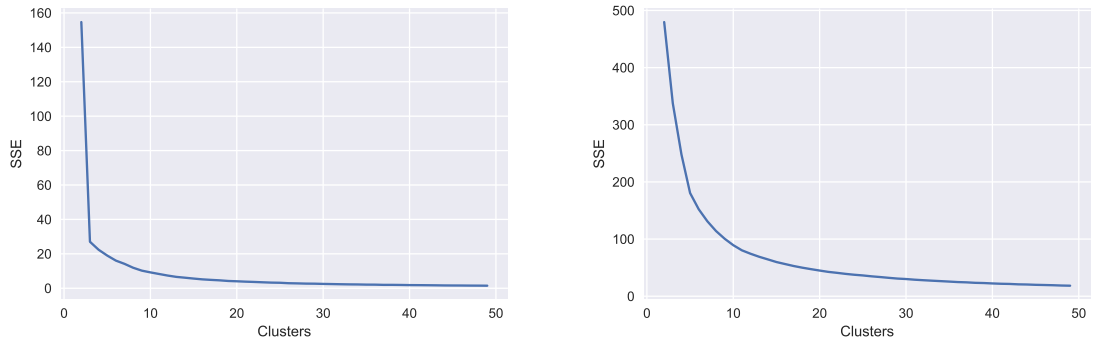
Approfondire su variabili discrete

Nell'implementazione dell'algoritmo da noi utilizzata è stato deciso di applicare la distanza euclidea come distance function.

2.1.2 Identification of the best value of k

Al fine di identificare il miglior numero k di clusters da utilizzare, abbiamo tenuto conto dell'Error Sum of Squares (SSE) per ogni iterazione dell'algoritmo, svolta a partire da un valore iniziale di k pari a 2 fino ad un valore massimo di 50. Rappresentato in Figura 2.1a troviamo l'andamento dell'SSE per i cluster relativi ai dipendenti che hanno lasciato l'azienda. Possiamo notare il valore ottimale di k pari a 3, che è la posizione sull'asse dei cluster dove la curva inizia il suo percorso discendente. In figura 2.1b possiamo invece vedere la stessa cosa, ma per i dipendenti che sono rimasti all'interno dell'azienda. In questo caso notiamo il valore ottimale di k pari a 5.

2.1.3 Characterization of the obtained clusters by using both analysis of the k centroids and comparison of the distribution of variables within the clusters and that in the whole dataset



(a) (b)

Figura 2.1: Prova

3 | Association Rules

4 | Classification