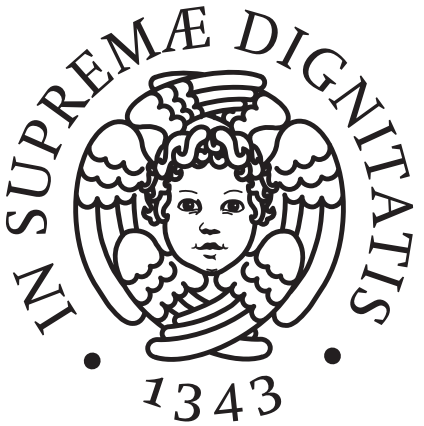


Università di Pisa

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

Capitolo 1

Data Understanding

1.1 Data semantics

In questo progetto viene analizzato il data set (simulato) **Human Resources Analytics** che contiene le informazioni sui dipendenti di un'azienda fittizia, suddivise in base alle variabili elencate nella Tabella 1.1.

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

Tabella 1.1: Variabili presenti nel Dataset e rispettivi tipi.

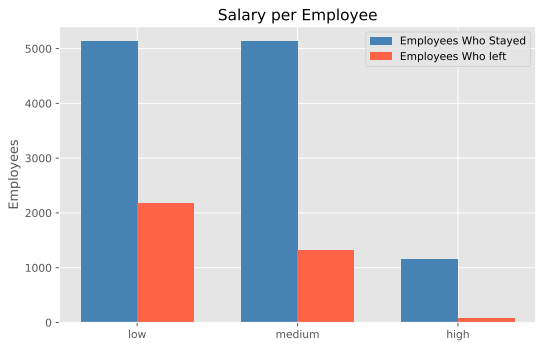
Con la prima variabile, *Satisfaction Level*, viene fornita una valutazione quantitativa del livello di soddisfazione di ciascun dipendente, in un range che va da un valore minimo di 0 ad un massimo di 1. *Last Evaluation* fornisce il tempo trascorso, in anni, dall'ultima valutazione delle performace 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.

1.2 Distribution of the variables and statistics

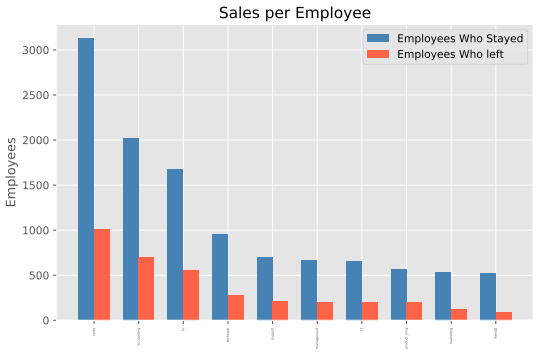
In questo paragrafo vengono presentati i grafici relativi alla distribuzione dei valori assunti dalle variabili descritte nella sezione precedente. Ogni figura viene presentata accompagnata da una descrizione della distribuzione rappresentata.

Nota sulle figure: i commenti vanno separati dalle captions facendoli a parte nel testo. Sistemare la grandezza delle annotazioni nei grafici

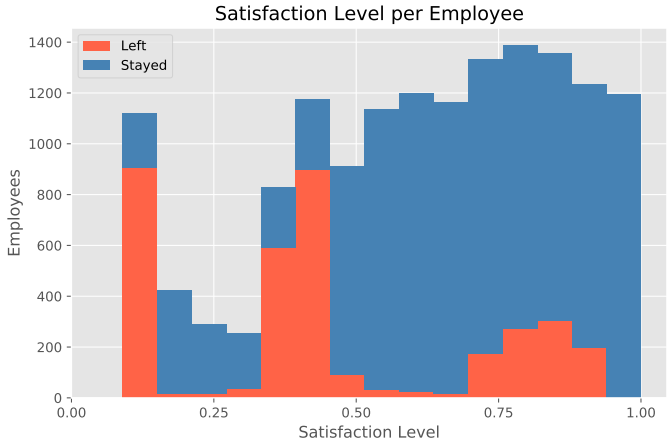
[1] https://www.kaggle.com/



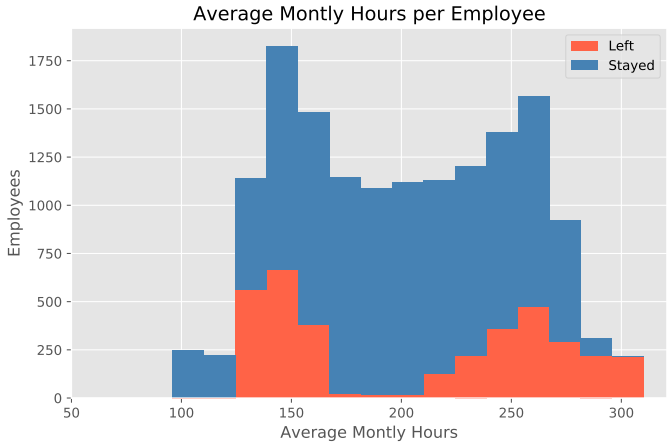
(a) Tra i 3571 impiegati che hanno lasciato l'azienda, 2172 rietrano nella categoria salariale *low*, mentre tra gli 11428 impiegati rimasti nell'azienda, 5144 rientrano nella categoria salariale *low*.



(b) Come possiamo osservare, il reparto *sales* è quello con il più alto numero di impiegati, sia che essi abbiano lasciato l'azienda sia che siano rimasti.

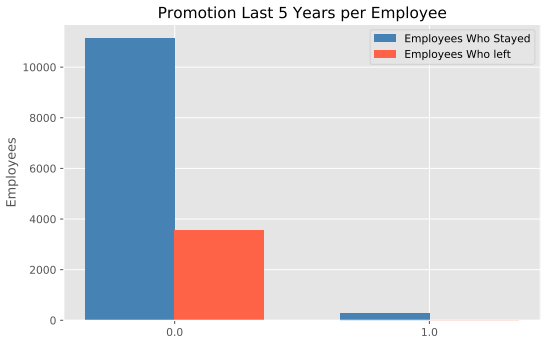


(c) Per gli impiegati che hanno lasciato l'azienda, la **media** del livello di soddisfazione è 0.44, la **deviazione standard** è 0.26, i valori **minimo** e **massimo** sono, rispettivamente, 0.09 e 0.92. Per quanto riguarda gli impiegati rimasti abbiamo invece una **media** pari a 0.67, una **deviazione standard** pari a 0.22 e un valore **minimo** e **massimo** pari a 0.12 e 1.0.

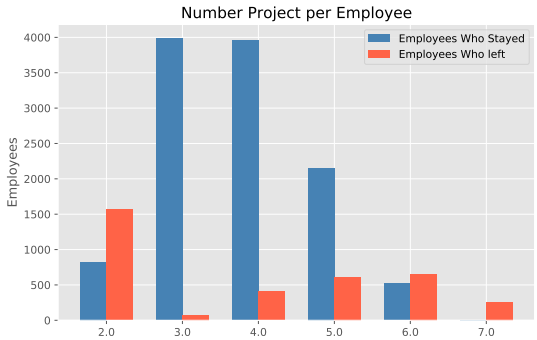


(d) Per gli impiegati che hanno lasciato l'azienda, la **media** delle ore di lavoro mensili è 207.42, la **deviazione standard** è 61.20, i valori **minimo** e **massimo** sono, rispettivamente, 126 e 310. Per quanto riguarda gli impiegati rimasti abbiamo invece una **media** pari a 199.06, una **deviazione standard** pari a 45.68 e un valore **minimo** e **massimo** pari a 96 e 287.

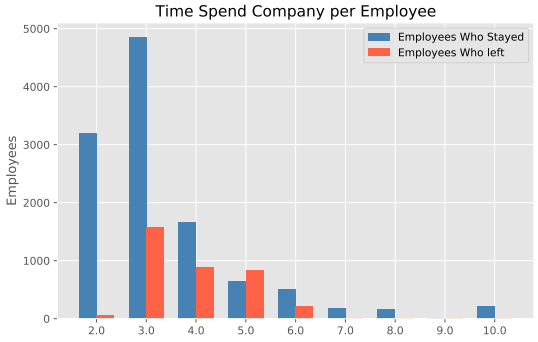
Figura 1.1: Da sinistra verso destra, e dall'alto verso il basso, sono raffigurate le distribuzioni delle variabili *Salary*, *Sales*, *Satisfaction Level* e *Average Montly Hours*. In rosso vengono rappresentati gli impiegati che sono rimasti nell'azienda, mentre in blu quelli che se ne sono andati. Per ogni variabile vengono fornite delle informazioni riguardo alla **media**, alla **deviazione standard** e ai valori **massimi** e **minimi** dei parametri analizzati.



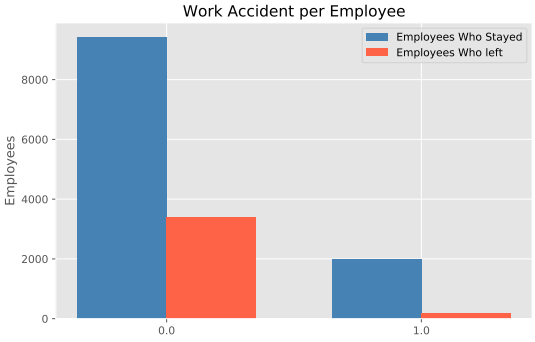
(a) Come possiamo osservare, la maggior parte degli impiegati che hanno lasciato l'azienda non ha avuto una promozione negli ultimi 5 anni. Soltato 19 impiegati su 3571 sono stati promossi. Lo stesso vale per gli impiegati rimasti, ma ovviamente in proporzioni diverse.



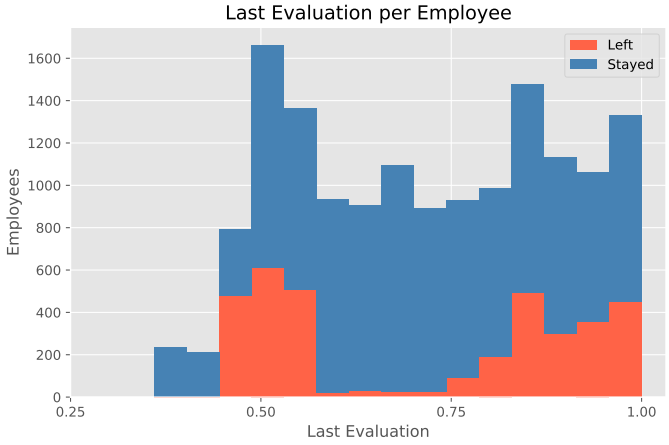
(b) La maggior parte degli impiegati che hanno lasciato l'azienda ha svolto soltato 2 progetti, mentre quelli che sono rimasti hanno svolto per la maggior parte 3 progetti, e in misura leggermente minore 4.



(c) Gli impiegati che se ne sono andati avevano trascorso per la maggior parte 3 anni all'interno dell'azienda, allo stesso modo gran parte degli impiegati al momento in azienda lavora al suo interno da 3 anni.



(d) Soltanto 169 impiegati tra i 3571 che se ne sono andati ha avuto un incidente sul lavoro, mentre gli impiegati ancora all'interno dell'azienda ad aver subito un incidente sono 2000 su 11428



(e)

Figura 1.2: Da sinistra verso destra, e dall'alto verso il basso, sono raffigurate le distribuzioni delle variabili *Promotion Last 5 Years*, *Number Project*, *Time Spend Company*, *Work Accident* e *Last Evaluation*. In rosso vengono rappresentati gli impiegati che sono rimasti nell'azienda, mentre in blu quelli che se ne sono andati. Per le prime quattro visualizzazioni sono fornite delle descrizioni sommarie, trattandosi di dati di tipo discreto.

1.3 Data quality

- Missing values

- Outliers

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

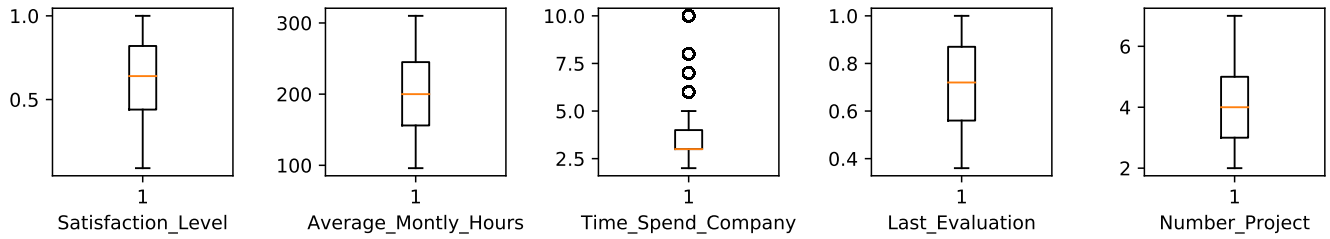


Figura 1.3: Boxplots per le variabili numeriche

1.4 Variable transformations

Analizzando il significato delle variabili presenti nel dataset, abbiamo deciso di rappresentare *Work Accident* e *Left* utilizzando il tipo categorico piuttosto che quello discreto. Questa scelta è stata motivata dall'analisi semantica delle due variabili, le quali forniscono una risposta del tipo "Sì o No" alle domande relative agli incidenti sul lavoro e all'abbandono o meno dell'azienda da parte dei dipendenti.

1.5 Pairwise correlations and eventual elimination of redundant variable

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc

dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

Capitolo 2

Clustering analysis

2.1 Clustering Analysis by K-means

2.1.1 Choice of attributes and distance function

2.1.2 Identification of the best value of k

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

Capitolo 3

Association Rules

Capitolo 4

Classification