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

Indice

Elenco delle figure

Capitolo 1

Data Understanding

1.1 Data semantics

Il progetto viene svolto sul dataset (simulato) **Human Resources Analytics**. Questo dataset contiene le informazioni sui dipendenti di un'azienda fittizia, suddivise in base ai seguenti campi:

Field	Type	Description
satisfaction_level	continuous	valutazione *quantitativa* del livello di soddisfazione di ciascun dipendente, ha un valore compreso tra 0 (minimo) e 1 (massimo)
last_evaluation	continuous	tempo trascorso, in anni, dall'ultima valutazione delle performace del dipendente
number_project	discrete	numero di progetti completati durante il periodo di lavoro
average_montly_hours	discrete	media delle ore di lavoro in un mese
time_spend_company	discrete	numero di anni trascorsi nell'azienda
work_accident	discrete	esprime con un 1 il coinvolgimento di un dipendente in un incidente sul lavoro, altrimenti viene impostato come 0
left	discrete	se il dipendente ha lasciato l'azienda viene impostato come 1, altrimenti come 0
promotion_last_5years	discrete	se il dipendente è stato promosso negli ultimi 5 anni viene impostato come 1, altrimenti come 0
sales	categorical	definisce il dipartimento nel quale il dipendente lavora
salary	ordinal	esprime il livello (*low*, *medium*, *high*), di salario nel quale rientra il dipendente

Per la tipizzazione dei campi presenti nel Dataset è stata seguita la classificazione fornita nel testo *Guide to Intelligent Data Analysis*, mentre, per la descrizione a parole, sono state seguite le direttive fornite nella pagina di Kaggle [1] nella quale è contenuto il Dataset [2].

[1] Kaggle
[2] Column Metadata

Figura 1.1: Distribuzione relativa alla colonna Salary.

1.2 Distribution of the variables and statistics

In questo paragrafo vengono presentati i grafici relativi alla distribuzione dei valori assunti dalle colonne descritte nella sezione precedente. Viene omesso il codice con i quali tali grafici sono stati costruiti. In Fig. 1.1 possiamo notare che la maggior parte degli impiegati che hanno deciso di lasciare l'azienda aveva un salario rientrate nella categoria *low*. Al contrario, un numero pressochè uguale di impiegati all'interno dell'azienda ha un salatio rientrate nelle categorie *low* e *medium*. La Fig. 1.2 rappresenta invece la distribuzione degli impiegati rispetto alla colonna Sales.

1.3 Assessing data quality

1.4 Variable transformations

1.5 Pairwise correlations and eventual elimination of redundant variable

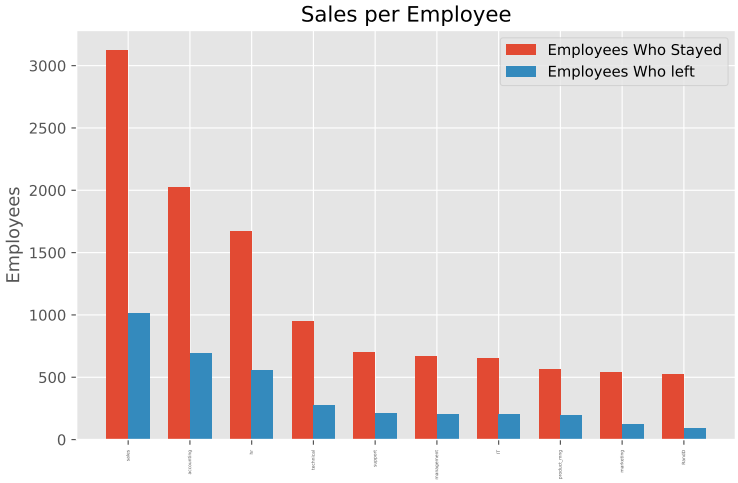

Figura 1.2: Distribuzione relativa alla colonna Sales

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