



UNIVERSITÀ DI PISA

PROGETTO PER IL CORSO DI DATA MINING

## **Analisi del Dataset Human Resources Analytics**

*Gianmarco Ricciarelli & Maria Cristina Uccheddu*

# Indice

<b>1 Data Understanding</b>	<b>4</b>
1.1 Data semantics . . . . .	4
1.2 Distribution of the variables and statistics . . . . .	5
1.3 Assessing data quality . . . . .	5
1.4 Variable transformations . . . . .	5
1.5 Pairwise correlations and eventual elimination od redundant variable . . . . .	5
<b>2 Clustering analysis</b>	<b>6</b>
2.1 Clustering Analysis by K-means . . . . .	6
2.1.1 Choice of attributes and distance function . . . . .	6
2.1.2 Identification of the best value of k . . . . .	6
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 . . . . .	6
<b>3 Association Rules</b>	<b>7</b>
<b>4 Classification</b>	<b>8</b>

# **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 <i>quantitativa</i> 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 performance del dipendente
number_project	continuous	numero di progetti completati durante il periodo di lavoro
average_montly_hours	continuous	media delle ore di lavoro in un mese
time_spend_company	continuous	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 ( <i>low, medium, high</i> ), 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<sup>1</sup> nella quale è contenuto il Dataset<sup>2</sup>.

---

<sup>1</sup>Kaggle

<sup>2</sup>Column Metadata

- 
- 1.2 Distribution of the variables and statistics**
  - 1.3 Assessing data quality**
  - 1.4 Variable transformations**
  - 1.5 Pairwise correlations and eventual elimination od redundant variable**

## **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**