TITOLO:RELAZIONE DI DATA MINING

Daniele Maria Di Nosse, Angelo Lasala, Raffaele Paradiso

21/11/2020

Indice

1 Introduzione **3**

2 Data Understanding **3**

2.1 Data Semantics 3

3 Clustering **5**

4 Conclusioni **5**

1 Introduzione

Determinare le possibili relazioni che intercorrono fra caratteristiche dei dipendenti di un'azienda può risultare di grande utilità per predire i possibili scenari lavorativi che posso verificarsi e gestire di conseguenza l'organizzazione del personale in maniera ottimale. Nel presente progetto ci si pone l'obiettivo di valutare tali legami tramite un approccio di data mining. Le informazioni che si sono utilizzate sono relative ad un data frame fittizio (leggermente modificato) generato da IBM e presente sul portale Kaggle(URL `https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset`). Non ci si è posto un obiettivo principale, ovvero la determinazione di legami, correlazioni e classificazioni relativi ad un singolo attributo rispetto a tutti gli altri, ma si è proceduto in maniera più generale ricoprendo uno spettro più ampio di possibili relazioni fra tutte le variabili.

Sebbene i dati a disposizione siano stati divisi in due sotto insiemi, uno di Train ed uno di Test, si è deciso di utilizzare l'intero insieme di records per tutti i tasks che non concernono algoritmi di Machine Learning

2 Data Understanding

2.1 Data Semantics

Nella prima fase dell'elaborazione si è studiato il data frame nella sua forma originale (Train + Test), valutando il numero degli attributi, la loro natura e dominio.

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EnvironmentSatisfaction	Gender	HourlyRate
0	27	Yes	Travel_Frequently	763	Sales	15	2	Medical	1	Male	59
1	30	Yes	Travel_Rarely	1079	Sales	16	4	Marketing	1	Male	70
2	56	No	Non-Travel	150	Research & Development	2	4	Technical Degree	4	Male	60
3	41	Yes		359	Human Resources	18	5	Human Resources	4	Male	89
4	42	No		642	Research & Development	1	3	Life Sciences	4	Male	76
5	42	No	Non-Travel	688	Sales	7	3	Life Sciences	3	Male	44
6	40	No	Travel_Frequently	684	Research & Development	6	3	Life Sciences	1	Female	51
7	54	No	Travel_Rarely	1302	Research & Development	6	4	Life Sciences	1	Female	80
8	45	No	Non-Travel	1402	Sales	28	4	Life Sciences	2	Female	98
9	37	No	Travel_Rarely	1282	Research & Development	5	3	Other	3	Male	58
10	36	No	Travel_Rarely	1381	Sales	4	4	Marketing	3	Female	72

	JobInvolvement	JobLevel	JobRole	JobSatisfaction	MaritalStatus	MonthlyIncome	MonthlyRate	umCompaniesWorke	Over18	OverTime	PercentSalaryHike
0	1	2	Sales Executive	4	Divorced	4298	22098	6	Y	Yes	14
1	3	3	Sales Executive	3	Married	5304	19002	2	Y	No	13
2	3	2	Manufacturing Direc	4	Divorced	6306	17433	2	Y	No	11
3	4	1	Human Resources	1	Married	6430	21495	0	Y	No	17
4	3	1	Research Scientist	4	Married	2766	21412	3	Y	No	22
5	2	3	Manager	4	Divorced	4332	25291	9		No	21
6	3	5	Research Director	3	Single	5605	6462	7	Y	No	13
7	4	2	Laboratory Technicia	1	Married	4440	19711	3	Y	Yes	13
8	2	1	Sales Representative	3	Married	8865	26204	0		No	23
9	3	5	Manager	3	Divorced		10735	4	Y	No	11
10	3	2	Sales Executive	3	Married	8008	12740	9	Y	No	15

	PerformanceRating	elationshipSatisfactio	StandardHours	StockOptionLevel	TotalWorkingYears	TrainingTimesLastYear	WorkLifeBalance	YearsAtCompany	YearsInCurrentRole	arsSinceLastPromoti	arsWithCurrManag
0	3	4	80	2	7	1	2	2	2	2	2
1	4	1		3	10	2	3	8	0	0	0
2	4	3	80	1	12	2	2	13	3	1	4
3	3	3	80	0	2	4	3	3	0	0	0
4	4	1		1	12	6	3	5	3	1	0
5	3	3	80	1	10	2	2	20	4	0	1
6	3	3	80	0	23	2	3	20	18	15	15
7	3	4	80	1	9	3	3	5	2	0	2
8	3	1	80	2	6	2	3		3	4	2
9	3	4		1	26	2	2	1	13	4	8
10	3	1	80	1	6	6	3		2	1	2

Figura 1: Primi 10 valori di tutti gli attributi

Come si può notare dalla tabella precedente, il numero di attributi è pari a 33. Si dividono in attributi numerici e categorici, ma ad uno sguardo più attento si nota che alcuni di essi, come, ad esempio, Education o Enviroment Satisfaction, presentano valori numerici che poco si adattano al loro significato. Si ha infatti che sussistono le seguenti uguaglianze

Education
1 : 'Below College'
2 : 'College'
3 : 'Bachelor'
4 : 'Master'
5 : 'Doctor'
EnvironmentSatisfaction
1 : 'Low'
2 : 'Medium'
3 : 'High'
4 : 'Very High'
JobInvolvement
1 : 'Low'
2 : 'Medium'
3 : 'High'
4 : 'Very High'
JobSatisfaction
1 : 'Low'
2 : 'Medium'
3 : 'High'
4 : 'Very High'
PerformanceRating
1 : 'Low'
2 : 'Good'
3 : 'Excellent'
4 : 'Outstanding'
RelationshipSatisfaction
1 : 'Low'
2 : 'Medium'
3 : 'High'
4 : 'Very High'
WorkLifeBalance
1 : 'Bad'
2 : 'Good'
3 : 'Better'
4 : 'Best'

Di conseguenza, il dominio di tali attributi è di tipo categorico od ordinale e non numerico. Organizzando tutte le variabili per la loro tipologia, si ottiene che

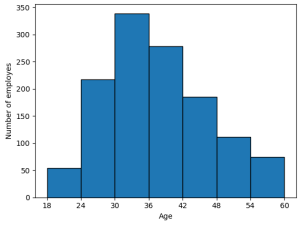
Per quanto riguarda il range di valori degli attributi risulta essere, come è possibile aspettarsi da quanto detto, molto più discretizzato per gli attributi ordinali che per gli attributi numerici. Inoltre, differisce molto da attributo ad attributo (anche di 4 ordini di grandezza), cosa che sottolinea sin da questo punto l'importanza di una trasformazione delle variabili.

Categorici : 8	Ordinali : 10	Numerici : 15
Attrition	Business Travel	Age
Department	Education	Daily Rate
Education Field	Enviroment Satisfaction	Distance From Home
Gender	Job Involvement	Hourly Rate
Job Role	Job Level	Monthly Income
Marital Status	Job Satisfaction	Monthly Rate
Over 18	Performance Rating	Num Companies Worked
Over Time	Relationship Satisfaction	Percent Salary Hike
	Stock Option Level	Standard Hours
	Work Life Balance	Total Working Years
		Training Time Last Year
		Years At Company
		Years In Current Role
		Years Since Last Promotion
		Years With Current Manager

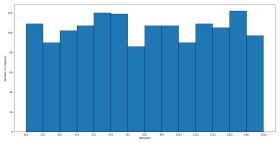
Figura 2: Classificazione degli attributi

3 Clustering

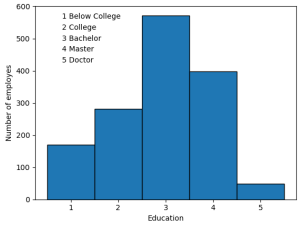
4 Conclusioni



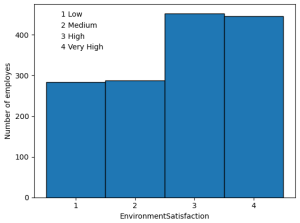
(a) Age



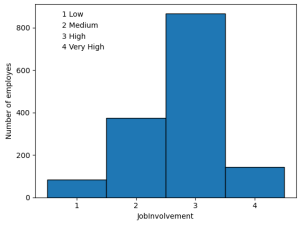
(b) Daily Rate



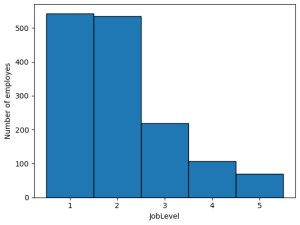
(d) Education



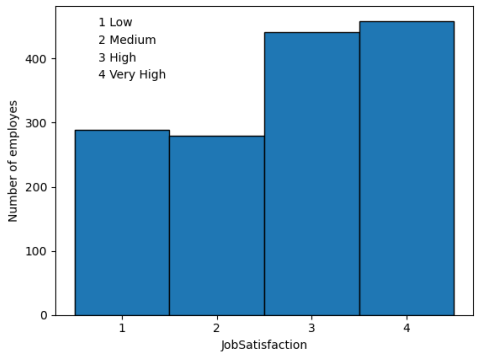
(e) Enviroment Satisfaction



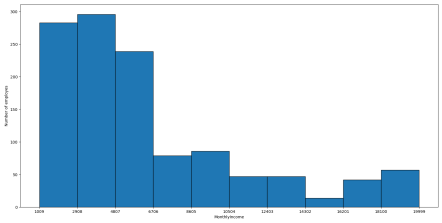
(g) Job Involvement



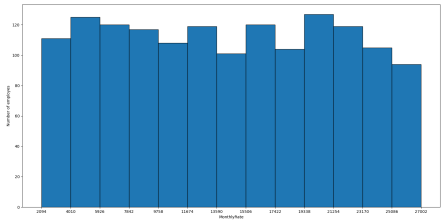
(h) Job Level



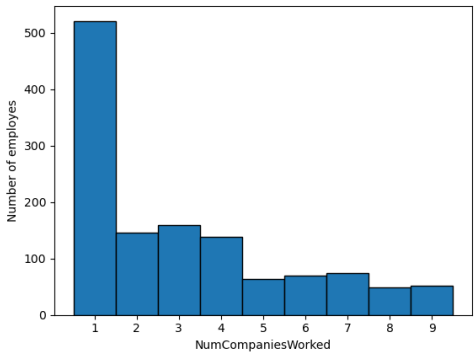
(a) Job Satisfaction



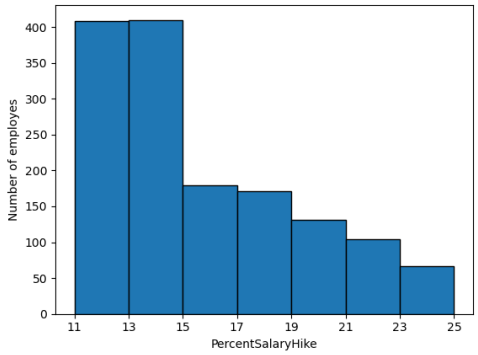
(b) Montlhy Income



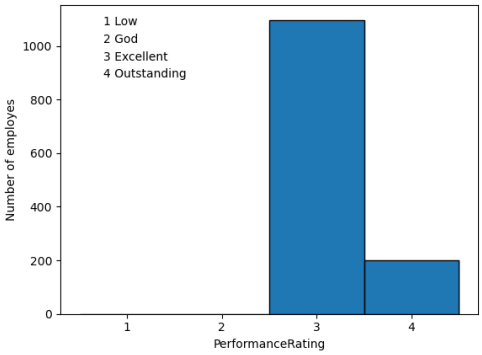
(c) MonthlyRate



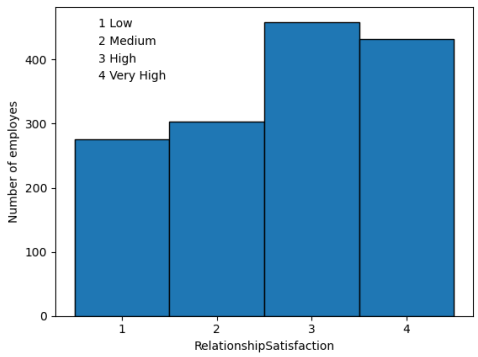
(d) Num Companies Worked



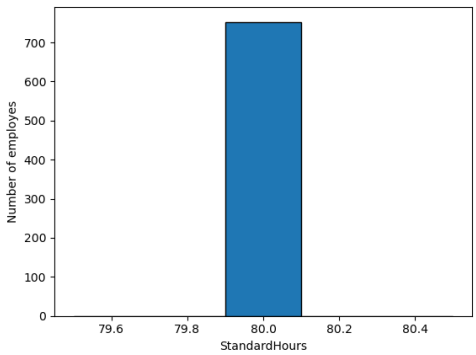
(e) Percent Salary Hike



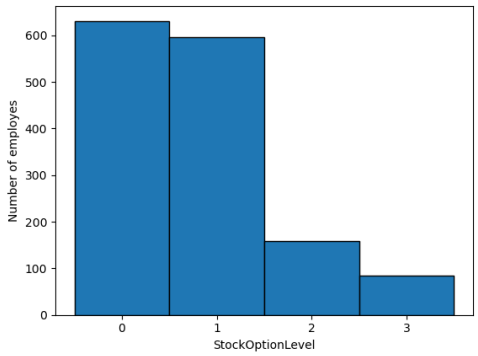
(f) Performance Rating



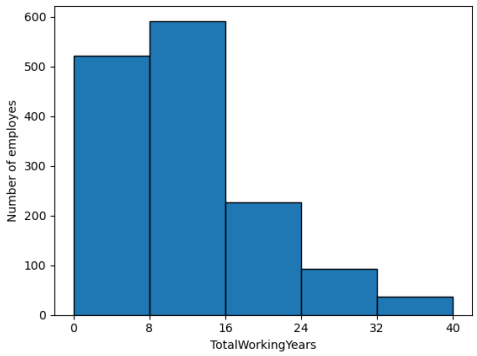
(g) Relationship Satisfaction



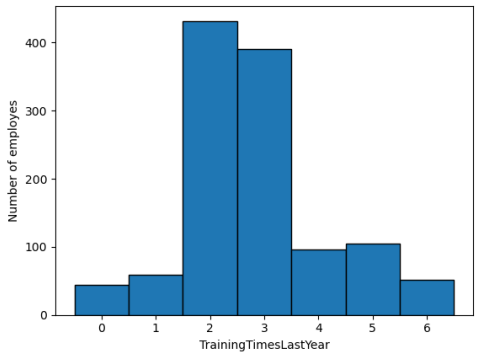
(h) Standard Hours



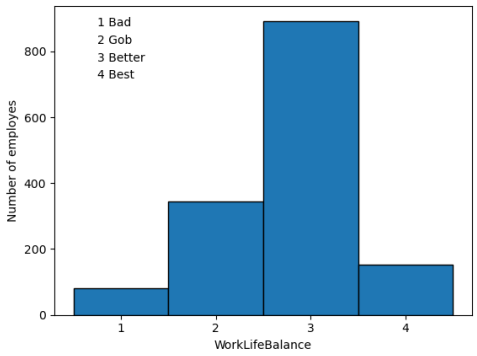
(a) Stock Option Level



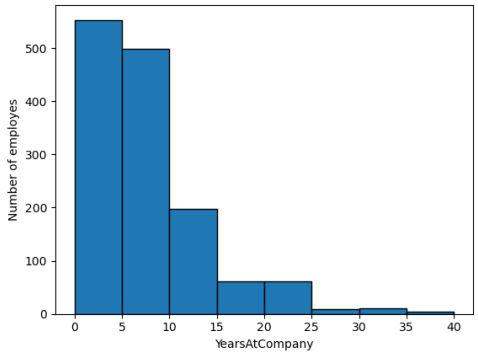
(b) Total Working Years



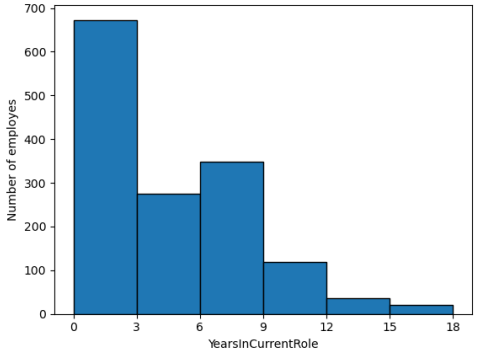
(c) Training Time Last Year



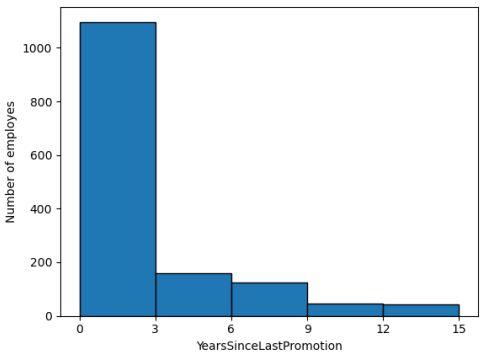
(d) Work Life Balance



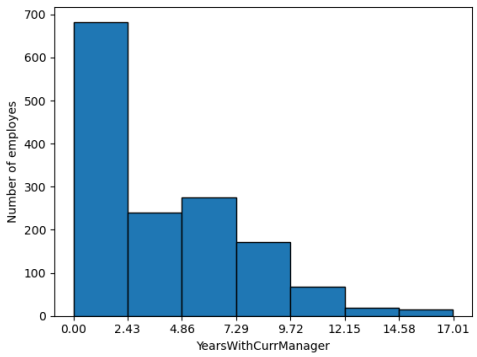
(e) Years At Company



(f) Years In Current Role



(g) Years Since Last Promotion



(h) Years With Current Manager