TITOLO:RELAZIONE DI DATA MINING

Daniele Maria Di Nosse, Angelo Lasala, Raffaele Paradiso

21/11/2020

Indice

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.

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	EnvironmentSatisfaction	JobInvolvement	JobSatisfaction
1 : 'Below College'	1 : 'Low'	1 : 'Low'	1 : 'Low'
2 : 'College'	2 : 'Medium'	2 : 'Medium'	2 : 'Medium'
3 : 'Bachelor'	3 : 'High'	3 : 'High'	3 : 'High'
4 : 'Master'	4 : 'Very High'	4 : 'Very High'	4 : 'Very High'
5 : 'Doctor'			

PerformanceRating	RelationshipSatisfaction	WorkLifeBalance
1 : 'Low'	1 : 'Low'	1 : 'Bad'
2 : 'Good'	2 : 'Medium'	2 : 'Good'
3 : 'Excellent'	3 : 'High'	3 : 'Better'
4 : 'Outstanding'	4 : 'Very High'	4 : 'Best'

Di conseguenza, il dominio di tali attributi è di tipo categorico od ordinale e non numerico(un attributo ordinale è effettivamente una sottocategoria categorica. Si è scelto comunque di elencarli separatamente). Inoltre, sebbene non si abbiano informazioni dettagliate sulle classi relative agli attributi JobLevel e StockOptionLevel, per la loro stessa natura si è deciso di trattarli come attributi ordinali. Organizzando tutte le variabili per la loro tipologia, si ottiene quindi che

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 1: Domini degli attributi

Per quanto riguarda il range di valori degli attributi risulta essere 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.

2.2 Analisi statistica

Le frequenze degli attributi categorici e le relative mode sono riportate nelle seguenti tabelle

Attriction	Educational Field	Departement	Gender	Over Time
'No': 83.9%	'Life Science': 41.2%	'Research and Development': 65.4%	'Male': 57.2%	'No': 71.7%
'Yes': 16.1%	'Medical': 31.6%	'Sales': 30.3%	'Female': 37.7%	'Yes': 28.3%
	'Marketing': 10.8%	'Human Resources': 4.3%	MISSING: 5.1%	
	'Technical Degree': 9.0%			
	'Other': 5.6%			
	'Human Resources': 1.8%			
Business Travel	**Job Rule**	**Matital Status**	**Over 18**	
'Travel Rarely': 64,4%	'Sales Executive': 22.2%	'Married': 45.8%	'Yes': 68.2%	
'Travel Frequently': 17,3%	'Research Scientist': 19.9%	'Single': 32,0%	MISSING: 31.8%	
'Non Travel': 9,4%	'Laboratory Technician': 17.6%	'Divorced':3 2,2%		
MISSING: 9,0%	'Manufacturing Derevtor': 9.9%			
	'Healthcare Representative': 8.9%			
	'Manager': 6.9%			
	'Sales Representative': 5.6%			
	'Research Director': 5.4%			
	'Human Resources': 3.5%			

Figura 2: Frequenze degli attributi categorici

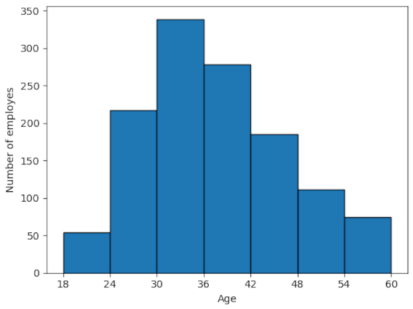
	Moda
Attrition	No
Educational Field	Life Science
Departement	Research and Development
Gender	Male
Over Time	No
Business Travel	Travel Rarely
Job Role	Sales Executive
Marital Status	Married
Over 18	Yes
Education	Bachelor
Enviroment Satisfaction	High
Job Involvement	High
Job Satisfaction	Very High
Performance Rating	Excellent
Relationship Satisfaction	High
Job Level	1
Work Life Balance	Better
Stock Option Level	0

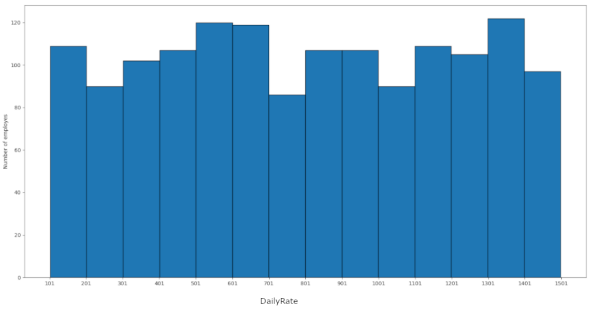
Figura 3: Mode

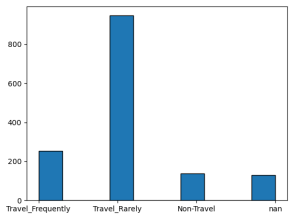
mentre le distribuzioni degli attributi ordinali e numerici con alcuni indici statistici sono rappresentate di seguito. Si può notare la forte asimmetria di molte distribuzioni ed un varianza molto grande in alcuni attributi. Tali problematiche dovranno essere sanate con opportune trasformazioni.

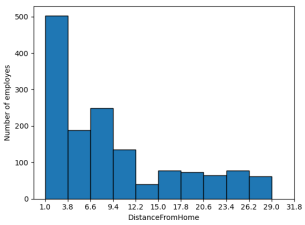
	Age	DailyRate	DistanceFromHome	HourlyRate	MonthlyIncome
count	1258	1470	1470	1470	1190
mean	37,11526232	802,4857143	9,192517007	65,89115646	6548,915966
std	9,068653862	403,5090999	8,106864436	20,32942759	4732,775331
min	18	102	1	30	1009
25%	30	465	2	48	2973,25
50%	36	802	7	66	4907,5
75%	43	1157	14	83,75	8437,5
max	60	1499	29	100	19999
	MonthlyRate	**NumCompaniesWorked**	**PercentSalaryHike**	**TotalWorkingYears**	**TrainingTimesLastYear**
count	1470	1470	1470	1470	1178
mean	14313,1034	2,693197279	15,20952381	11,27959184	2,810696095
std	7117,786044	2,498009006	3,659937717	7,780781676	1,302499143
min	2094	0	11	0	0
25%	8047	1	12	6	2
50%	14235,5	2	14	10	3
75%	20461,5	4	18	15	3
max	26999	9	25	40	6
	YearsAtCompany	**YearsInCurrentRole**	**YearsSinceLastPromotion**	**YearsWithCurrManager**	**StandardHours**
count	1396	1470	1470	1470	753
mean	6,94269341	4,229251701	2,187755102	4,123129252	80
std	6,033444155	3,623137035	3,222430279	3,568136121	0
min	0	0	0	0	80
25%	3	2	0	2	80
50%	5	3	1	3	80
75%	9	7	3	7	80
max	40	18	15	17	80

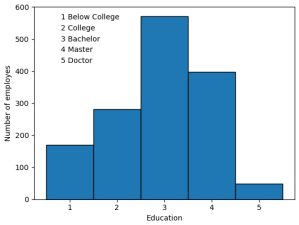
Figura 4: Indici statistici per gli attributi numerici

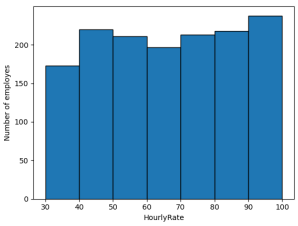

Number of employes
Age

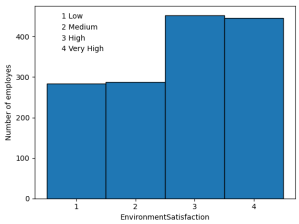

Number of employes
DailyRate

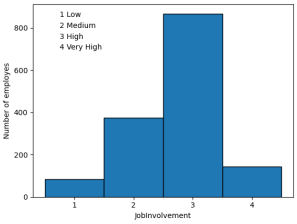

Travel_Frequently
Travel_Rarely
Non-Travel
nan

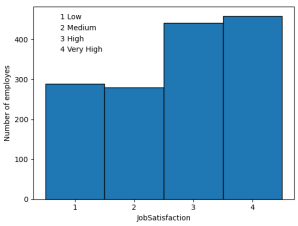

Number of employes
DistanceFromHome

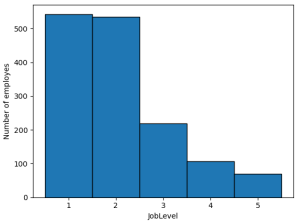

1 Below College
2 College
3 Bachelor
4 Master
5 Doctor
Number of employes
Education

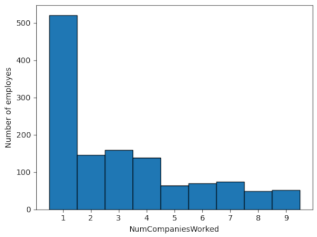

Number of employes
HourlyRate

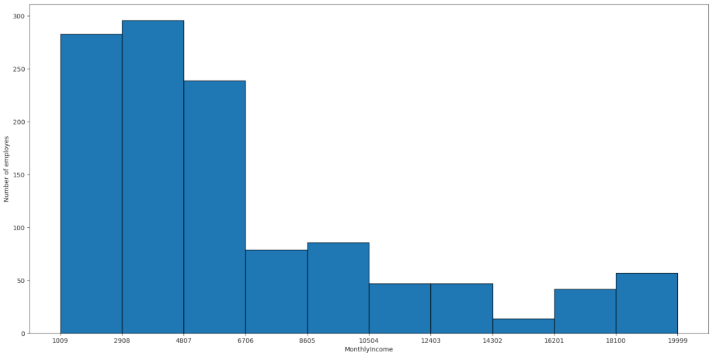

1 Low
2 Medium
3 High
4 Very High
Number of employes
EnvironmentSatisfaction

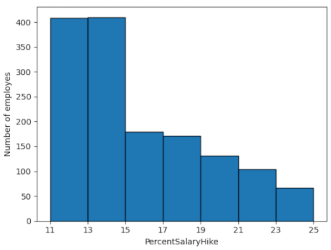

1 Low
2 Medium
3 High
4 Very High
Number of employes
JobInvolvement


1 Low
2 Medium
3 High
4 Very High
Number of employes
JobSatisfaction


Number of employes
JobLevel


Number of employes
NumCompaniesWorked


Number of employes
MonthlyIncome


Number of employes
PercentSalaryHike

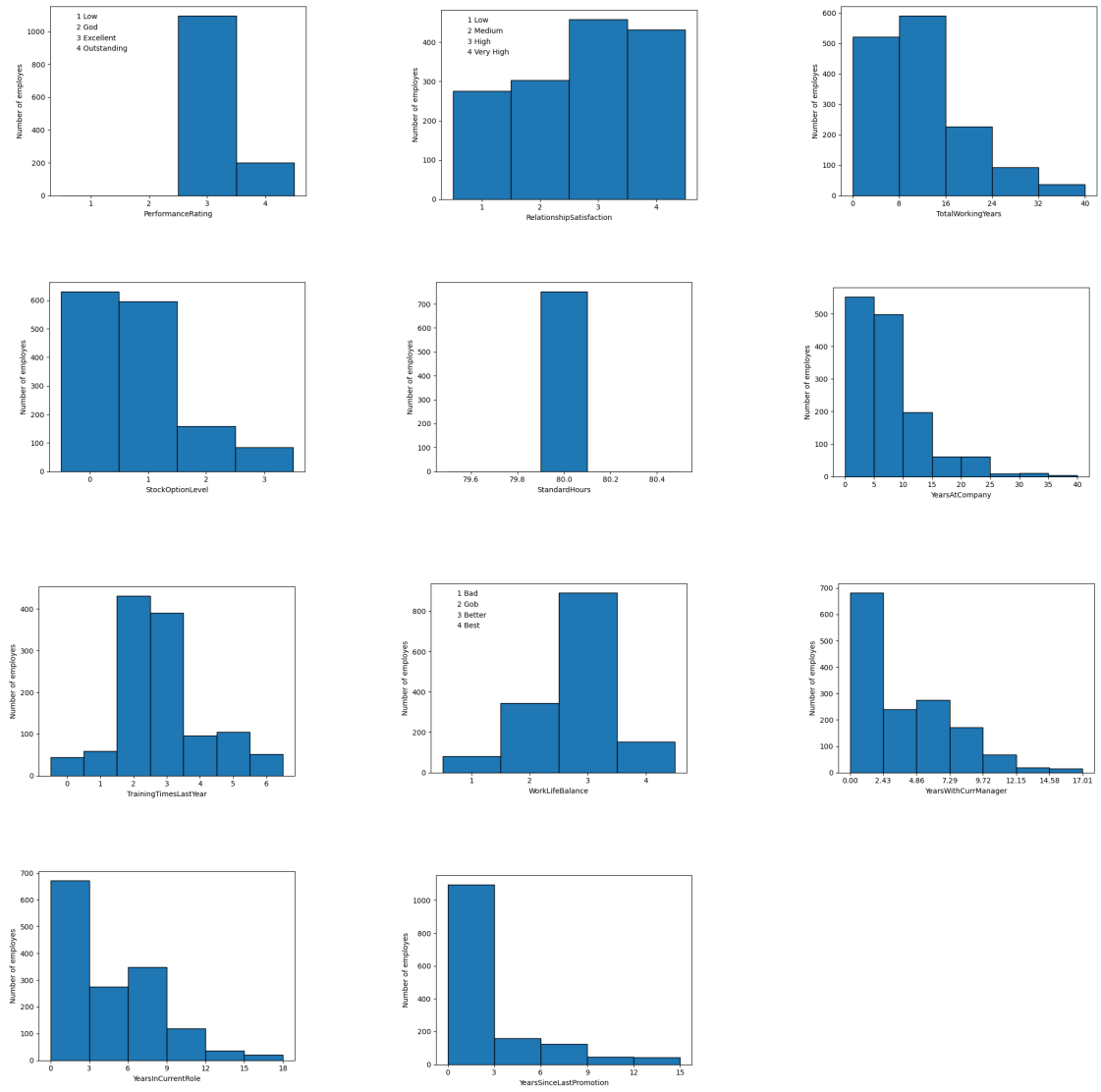


Figura 5: Istogrammi attributi numerici ed ordinali

2.3 Data Quality : Outliers e Missing values

La qualità dei dati è fortemente influenzata, negativamente, dalla presenza di outliers e di missing values. Algoritmi di clustering e correlazioni fra gli attributi possono restituire risultati falsificati se non si gestiscono in maniera appropriata tali valori. Nel data frame utilizzato la loro presenza è evidente, infatti

	Count Missing Values
Age	212
BusinessTravel	131
Gender	75
MonthlyIncome	280
Over18	468
PerformanceRating	172
StandardHours	717
TrainingTimesLastYear	292
YearsAtCompany	74

Figura 6: Count dei missing values

Gli attributi StandardHours ed Over18 presentano una qualità molto scarsa: nel primo circa la metà dei records sono mancanti e la restante parte ha un unico valore, mentre il secondo, oltre a contenere anch'esso una quantità significativa di missing values, non rappresenta in ogni caso un attributo di grande importanza, considerando che

la stragrande maggioranza dei dipendenti di un'azienda sono maggiorenni. Per tali motivi, si è deciso di eliminare questi due attributi.

Per la determinazione degli outliers si sono valutati sia test puramente statistici (Grubbs's test) che metodi di visualizzazione (Box Plot, Principal Component Analysis e scatter plot). Come è noto, per utilizzare approcci del primo tipo bisogna fare delle assunzioni sulla distribuzione sottostante dei valori esaminati. In particolare, il Grubbs's test, applicabile singolarmente agli attributi, richiede che i dati siano distribuiti normalmente, cosa non vera in questo caso. Di conseguenza, tale metodo è stato scartato. Il Principal Component Analysis, d'altro canto, è uno dei metodi maggiormente utilizzati nella ricerca di outliers in situazioni alto-dimensionali. Proiettando lo spazio n-dimensionale in uno spazio q-dimensionale ($q < n$) costruito tramite i vettori normalizzati della matrice di correlazione, si cerca di mantenere il più intatta possibile la varianza negli attributi. Nel caso in esame, la frazione di varianza conservata non risulta essere significativa (circa 0.4), inficiando inevitabilmente i risultati ottenuti. Anche la visualizzazione degli scatter plot confrontati con gli attributi categorici non ha evidenziato alcun punto identificabile con un outlier. L'unico metodo che ha avuto successo per la determinazione degli outliers è stata la visualizzazione dei Box Plot per i singoli attributi. Si è proceduto quindi alla loro rimozione tramite eliminazione delle righe corrispondenti.

3 Data Preparation

In questa fase del lavoro ci si è posto l'obiettivo di trasformare e preparare il set di dati all'analisi successiva. I problemi precedentemente evidenziati sono stati qui risolti.

Come primo task sono stati gestiti i missing values. Nell'attributo BusinessTravel presenta una frequenza di NaN pari al circa 9%, confrontabile con le frequenze degli altri valori . Siccome la granulosità dell'attributo ricopre in maniera completa lo spettro delle classi plausibilmente ad esso associabili, si è deciso di valutare se ci fosse dipendenza con gli altri attributi presenti nel data frame. Per quanto riguarda quella con gli numerici, sono stati utilizzati gli scatter plot, mentre per quelli nominali è stato eseguito il test di indipendenza del chi quadro. In entrambi i casi non si sono evinte dipendenze significative ($pvalue$>0.05 sempre). Di conseguenza tale attributo è stato scartato.
Per quanto riguarda PerformanceRating, si è aggiunta una nuova classe 'MISSING', poiché si è notato che la granulosità dell'attributo non ricopre tutto lo spettro plausibile. Si presuppone che i valori MISSING possano appartenere ad una classe di ordine inferiore ad Excellent.
Queste due considerazioni non sono applicabile all'attributo Gender per il quale si è scelto semplicemente di sostituire ai missing values valori estratti dalla distibuzione nota.
Procedimento analogo è stato applicato a tutti gli attributi numerici che presentano valori mancanti, l'unica differenza è che in questo caso i valori sostitutivi sono le medie degli intervalli dei bins degli istogrammi.
Come secondo task sono stati valutati gli outliers.
Il metodo di visualizzazione grafica dei Box Plot evidenzia la presenza di outliers solo in tre attributi numerici: TrainingTimeLastYear, TotalWorkingYears, YearsAtCompany

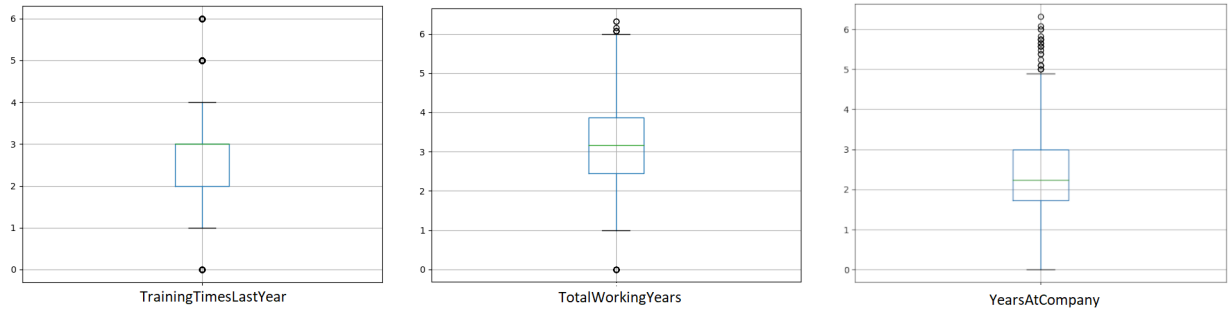


Figura 7: Box Plot degli attributi che presentano outliers

Il data frame dopo questa prima preparazione risulta contienere il 36% di dati in meno rispetto a quello di partenza. Funzioni di trasformazione sono state applicate ad attributi numerici con lo scopo di rimediare ad alcune caratteristiche delle loro distribuzioni quali l'asimettrie e un valore spropositato della deviazione standar. In particolre è stata applicata la radice quadrata ad DistanceFromHome (skew da 0.95 a 0.40), NumCompaniesWorked (skew da 1.03 a 0.03), PercentSalaryHike (skew da 0.82 a 0.65), TotalWorkingYears (skew da 1.12 a 0.18), YearsAtCompany (skew da 1.76 a 0.43), YearsInCurrentRole(skew da 0.92 a -0.25), YearsSinceLastPromotion (skew da 1.98 a 0.74) e YearsWithCurrManager (skew da 0.83 a -0.25); invece ad MonthlyIncome è stato applicato il logaritmo naturale (varianza da 4710 a 0.67) . Di seguito sono riportate alcune distribuzioni delle variabili trasformate.

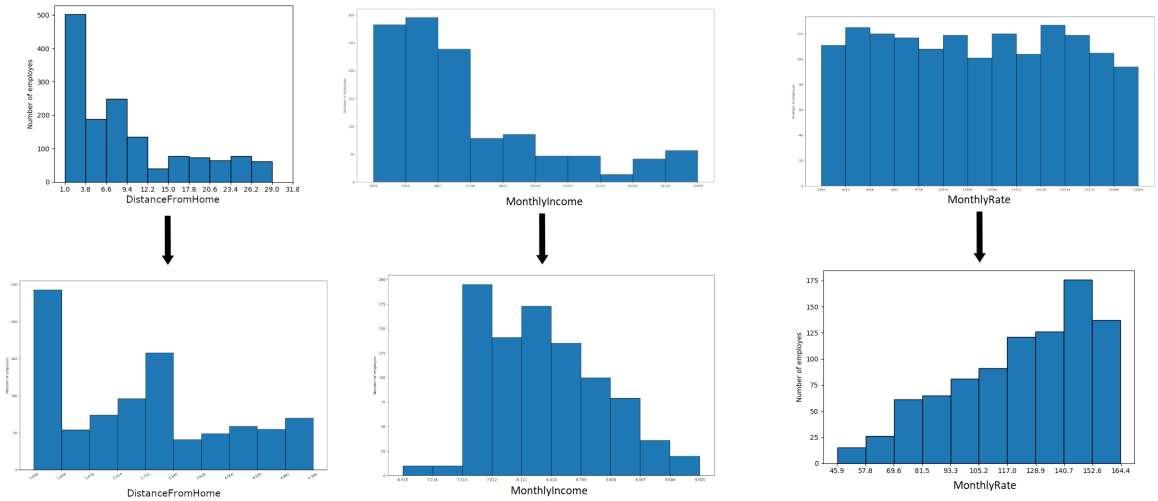


Figura 8: Alcune delle trasformazioni effettuate

Come terzo task sono stati eliminati ed aggiunti nuovi attributi.
In luogo di TotalWorkingYears e YearsAtCompany si è scelto di introdurre il loro rapporto, denominato FractionAtCompany che rappresenta la frazione di anni lavorativi del singolo dipendente nell'azienda; analogamente si è proceduto per MonthlyIncome e MonthlyRate sostituiti da RateIncome, indice di quanto l'azienda spende per un impegato in rapporto al suo stipendio. Inoltre, siccome DailyRate e HourlyRate contengono la stessa informazione di MonthlyRate, si sono eliminati.
Inoltre, YearsInCurrentRole, YearInCurrManager e YearsSinceLastPromotion sono caratterizzati da una correlazione significativa e quindi si è deciso di mantenere solamente YearsInCurrentRole ell'analisi a seguire.

Figura 9: Correlazione tra YearsInCurrentRole, YearInCurrManager e YearsSinceLastPromotion

Infine è stata calcolata la matrice di correlazione lineare fra gli attributi numerici e i valori del p value ottenuti tramite test del chi quadro per l'interdipendenza fra gli attributi categorici.

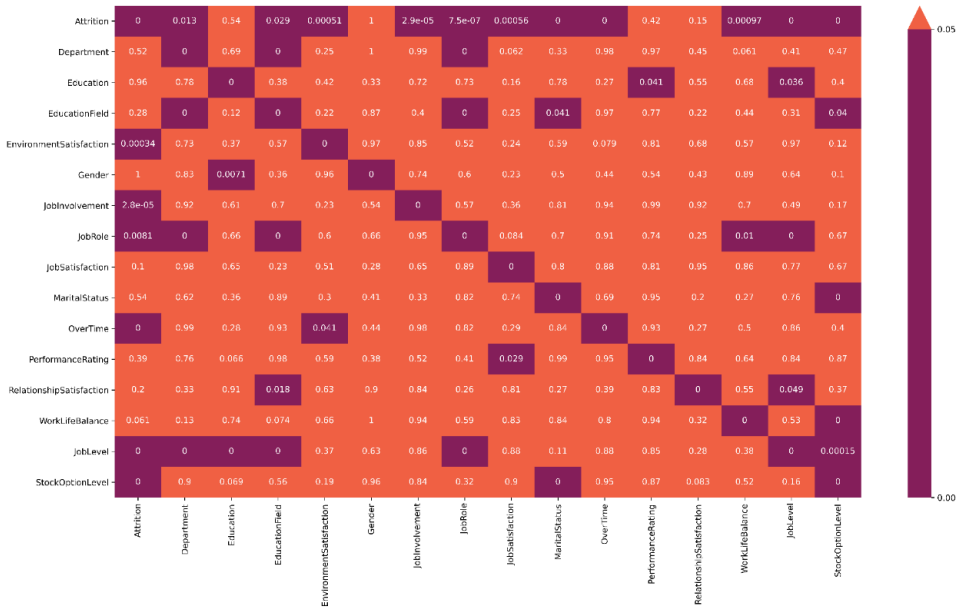


Figura 10: Matrice dei p value per gli attributi categorici

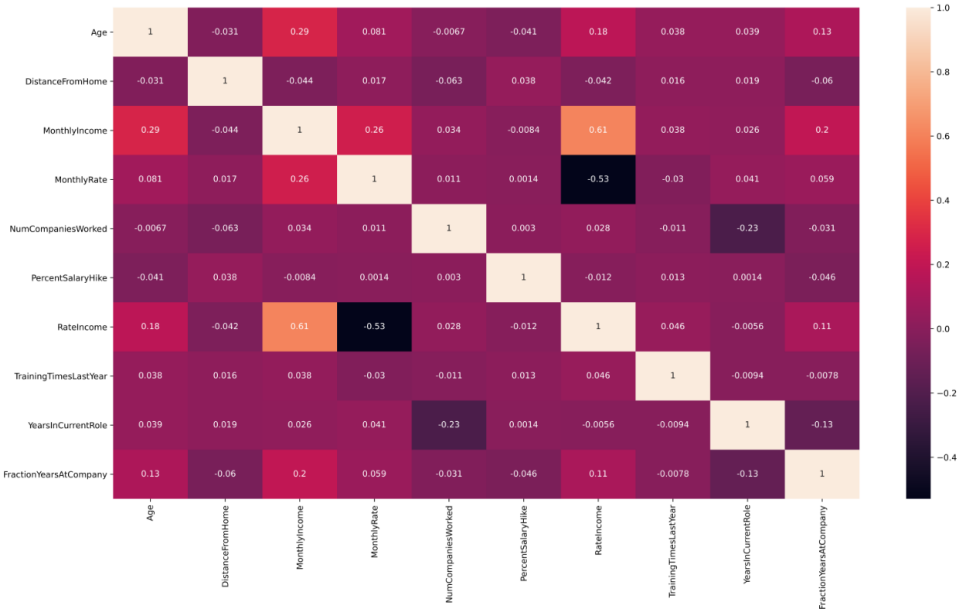


Figura 11: Matrice dei p value per gli attributi categorici

4 Clustering

Preparato il data frame si è proceduto all'analisi degli algoritmi di clustering: K-Means e DB Scan. La dimensionalità del data frame (10) è stata considerata troppo elevata per ottenere risultati consistenti, quindi sono stati indagati sottoinsiemi 4-5 dimensionali di attributi alla ricerca un qualche tipo di clusterizzazione. Come metrica è stata usata la distanza euclidea.
Si vuole precisare che per la visualizzazione dei clusters è stato utilizzato uno spazio tridimensionale poichè, soprattutto nel K-Means, una visualizzazione bidimensionale portava ad un mixing eccesivo dei cluster stessi.

4.1 K-Means

Per quanto riguarda il K-Means i sottoinsiemi che hanno mostrato i risultati migliori sono:

1. PercentSalaryHike, FractionYearsAtCompany, YearsInCurrentRole, RateIncome, NumCompaniesWorked

2. DistanceFromHome, FractionYearsAtCompany, RateIncome, YearsInCurrentRole, TrainingTimesLastYear(2)
3. PercentSalaryHike, DistanceFromHome, RateIncome, YearsInCurrentRole, NumCompaniesWorked(4)
4. DistanceFromHome, FractionYearsAtCompany, TrainingTimesLastYear, PercentSalaryHike, YearsInCurrentRole(6)

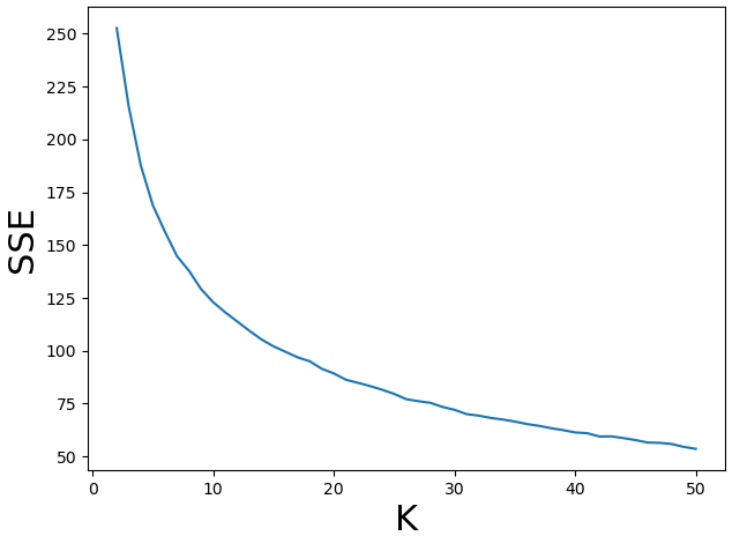


Figura 12: SSE in funzione di K

La scelta del numero di cluster K è stata presa osservando l'andamento del SSE in funzione di K (Figura 12), simile in tutti e quattro i casi esaminati; con lo scopo di aver un buon compromesso fra i due l'algoritmo è stato eseguito per K uguale a 3, 4 e 5. Dai risultati ottenuti si evince che, sebbene con $K = 5$ il valore delle SSE è minore rispetto agli altri due casi, non si apprezzano cluster evidenti: ve ne sono sempre due eccessivamente mescolati. Con $K = 3$ la divisione fra i clusters è sicuramente ben evidente ma, con $K = 4$, si ottengono comunque buoni risultati con il vantaggio di un SSE minore.
Nei seguenti grafici sono mostrati i risultati ottenuti inoltre, per redendere più chiara la posizione dei centroidi, sono riportate anche le loro cordinate organizzate in parallelo per ciascun sottoinsieme usato.

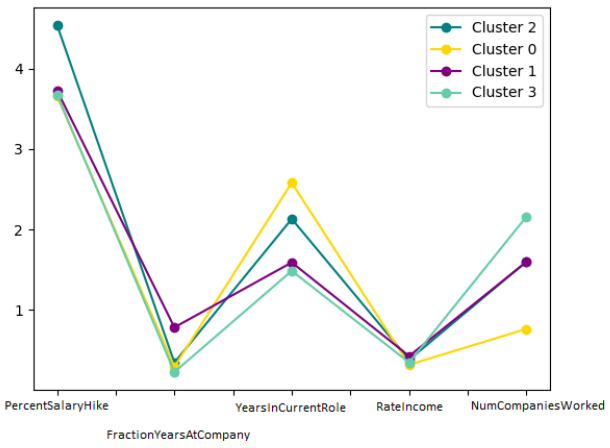


Figura 13: Parallel coordinates dei centroidi

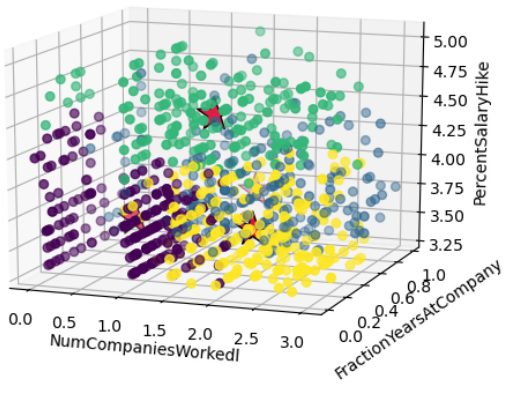


Figura 14: Cluster sottoinsime 3

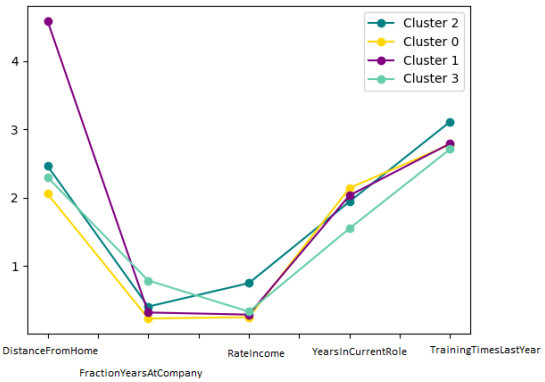


Figura 15: Parallel coordinates dei centroidi

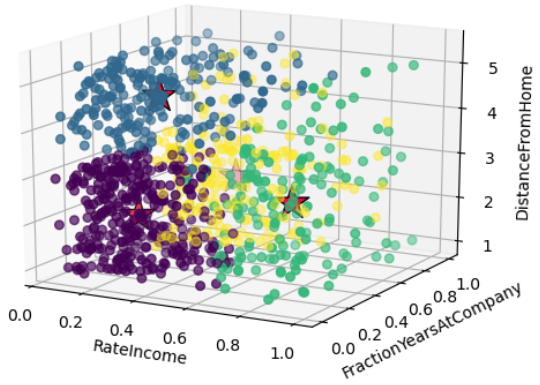


Figura 16: Cluster sottoinsime 2

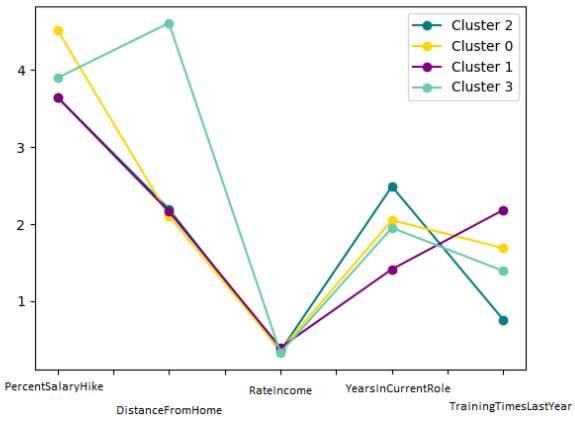


Figura 17: Parallel coordinates dei centroidi

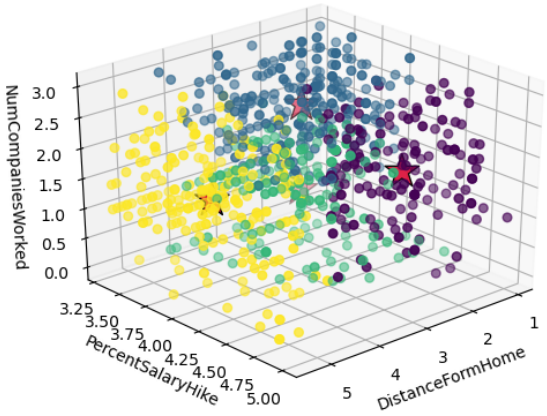


Figura 18: Cluster sottoinsime 3

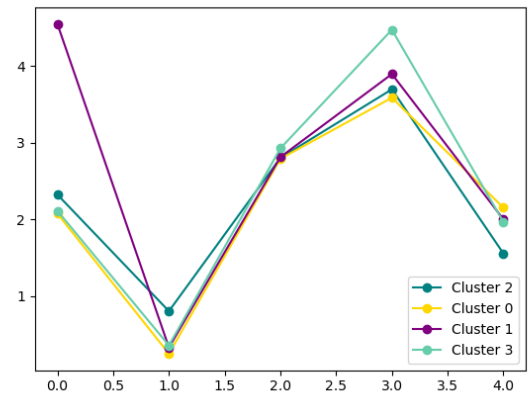


Figura 19: Parallel coordinates dei centroidi

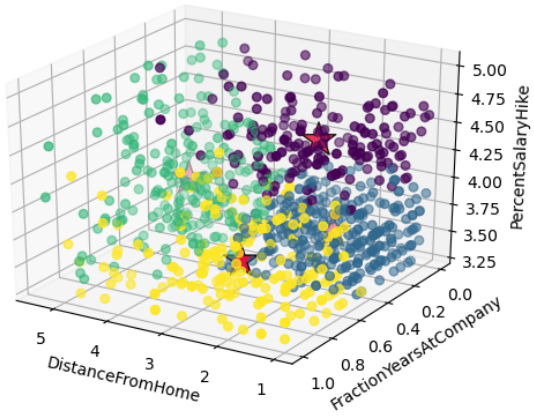


Figura 20: Cluster sottoinsime 4

4.2 DB-Scan

5 Conclusioni