

Progetto Data Mining

Andrea Carnevale
Federico Canepuzzi

IBM HR Analytics Employee Attrition & Performance

Data Science & Business Informatics
Università di Pisa
Anno Accademico 2020/2021

Contents

1	Introduzione	1
2	Data Understanding	1
2.1	Data semantics	1
2.2	Distribuzione delle variabili e analisi statistiche	1
2.3	Data quality	4
2.3.1	Errori sintattici e semantici	5
2.3.2	Missing Values	5
2.3.3	Outliers	6
2.4	Correlazione ed eliminazione di variabili	6
2.5	Trasformazione delle variabili	7
3	Clustering	8
3.1	Scelta degli attributi	8
3.2	K-Means	8
3.2.1	Identificazione del migliore valore di K	8
3.2.2	Caratterizzazione dei cluster ottenuti	9
3.3	DBScan	10
3.3.1	Studio dei parametri per il clustering	10
3.3.2	Caratterizzazione e interpretazione dei cluster ottenuti	10
3.4	Hierarchical	11
3.4.1	Visualizzazione e discussione di dendrogrammi differenti ottenuti con algoritmi differenti	11
3.5	Valutazione finale del miglior approccio di clustering	12
4	Classification	12
4.1	Decision Tree	12
4.1.1	Introduzione	12
4.1.2	Modello 1	13
4.1.3	Modello 2	13
4.1.4	Modello 3	14
4.1.5	Scelta e valutazione sul test set del miglior modello	14
4.2	KNN	15
4.2.1	Introduzione	15
4.2.2	Selezione dei parametri	16
4.2.3	Valutazione sul test set	17
4.3	Discussione sul miglior modello	17
5	Association Rules Mining	17
5.1	Data preparation	17
5.2	Frequent Patterns	18
5.3	Association Rules	19
5.3.1	Sostituzione Missing values	20
5.3.2	Predizione dell'Attrition	20

1 Introduzione

Il progetto si propone di analizzare il dataset IBM HR al fine di individuare le possibili cause che provocano attrition nell'azienda.

Con il termine inglese attrition si intende la graduale riduzione di personale che si verifica quando gli impiegati si dimettono e non vengono sostituiti. Nel dataset sono presenti 1176 record con 33 attributi ciascuno.

2 Data Understanding

2.1 Data semantics

In questa sezione vengono riportati in tabella gli attributi con la loro descrizione e la loro tipologia.

Nome	Descrizione	Tipo
Age	Età dell'impiegato	Numerico discreto
Attrition	Indica se l'impiegato presenta o meno attrition	Categorico
BusinessTravel	Frequenza dei viaggi di lavoro	Categorico
*Rate	Costo dell'impiegato dal punto di vista dell'azienda	Numerico continuo
MonthlyIncome	Paga mensile dell'impiegato	Numerico continuo
Department	Dipartimento in cui lavora l'impiegato	Categorico
DistanceFromHome	Distanza casa-lavoro	Numerico discreto
Education	Titolo di studio $\in \{1,..,5\}$ in ordine crescente di grado	Numerico discreto
EducationalField	Campo del percorso di studi	Categorico
Gender	Sesso dell'impiegato	Categorico
JobLevel	Livello di professione $\in \{1,..,5\}$ in ordine crescente di livello	Numerico discreto
JobRole	Ruolo di lavoro nell'azienda	Categorico
MaritalStatus	Stato civile dell'impiegato	Categorico
*Satisfaction	Esprimono il grado di soddisfazione dell'impiegato nei diversi aspetti $\in \{1,..,4\}$ in ordine crescente di soddisfazione	Numerico discreto
JobInvolvement	Esprime il grado di coinvolgimento nel lavoro attuale $\in \{1,..,4\}$	Numerico discreto
WorkLifeBalance	Esprime il grado di bilancio nella vita lavorativa $\in \{1,..,4\}$	Numerico discreto
NumCompaniesWorked	Numero di compagnie in cui l'impiegato ha lavorato	Numerico discreto
Over18	Indica se l'impiegato è maggiorenne o meno	Categorico
OverTime	Indica se l'impiegato ha svolto straordinari	Categorico
PercentSalaryHike	Percentuale di aumento dello stipendio in caso di promozione	Numerico discreto
PerformanceRating	Esprime una valutazione sulle performance dell'impiegato $\in \{1,..,4\}$	Numerico discreto
StandardHours	Ore lavorative	Numerico discreto
StockOptionLevel	Livello delle Stock Option $\in \{0,..,3\}$	Numerico discreto
*Years	Anni nello stesso ruolo, con lo stesso manager ecc..	Numerico discreto

Table 1: Descrizione degli attributi

2.2 Distribuzione delle variabili e analisi statistiche

In questa sezione verrà studiato il dataset, riportando analisi statistiche e informazioni sulla distribuzione degli attributi.

La prima parte avrà un carattere del tutto generale in modo da dare un'ampia visione del dataset, per poi spostare il nostro focus sull'attrition, e come possa essere influenzato dagli altri attributi.

Nell'azienda vi è un rapporto bilanciato tra impiegati uomini (59,4%) ed impiegate donne (40,6%) (Figura 1a), con un'età che varia dai 18 ai 60 anni con dei picchi tra i 28 e i 42 anni (Figura 1c). Non ci sono grandi differenze di guadagno mensile in base al sesso, con la media maschile pari a 6487 euro e quella femminile pari a 6709 euro. Il 46.1% degli individui è sposato, il 32.6% è single e il 21.3% è divorziato (Figura 1b).

Per quanto riguarda l'istruzione del personale si osserva dalla figura 2a che il 39% ha ottenuto il Bachelor Degree[3], il 26,5% il Master Degree[4], mentre sono in minoranza le altre categorie

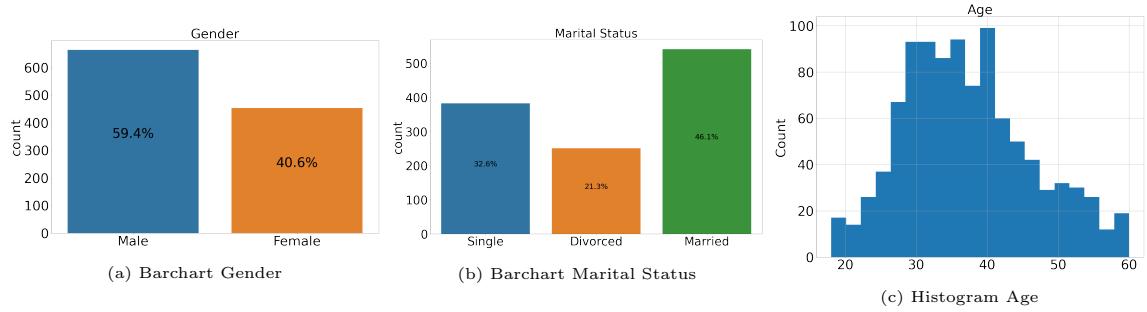


Figure 1

(Below College[1], College[2], Doctor[5]). Si può vedere dalla Figura 2b che la maggior parte degli individui proviene da campi di studio in Life Science o Medical.

Nei grafici sottostanti sono rappresentati i vari ruoli di lavoro presenti in azienda (Figura 3a), con associati i relativi MonthlyIncome (Figura 3b). Research Director (7611.7) è il lavoro con la paga media più alta mentre il lavoro di Manager (6013.5) ha la più bassa. Si è notato inoltre che l'attributo MonthlyIncome sia strettamente correlato agli anni dell'individuo, in particolare al crescere degli anni aumenta il salario medio.

I lavori più diffusi in azienda sono Sales Executive(22,2%), Research Scientist(20,4%) e Laboratory Technician (18,0%) (Figura 3a).

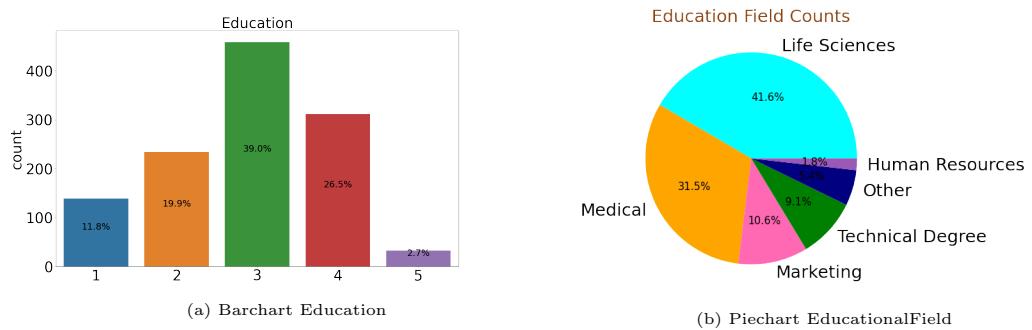


Figure 2

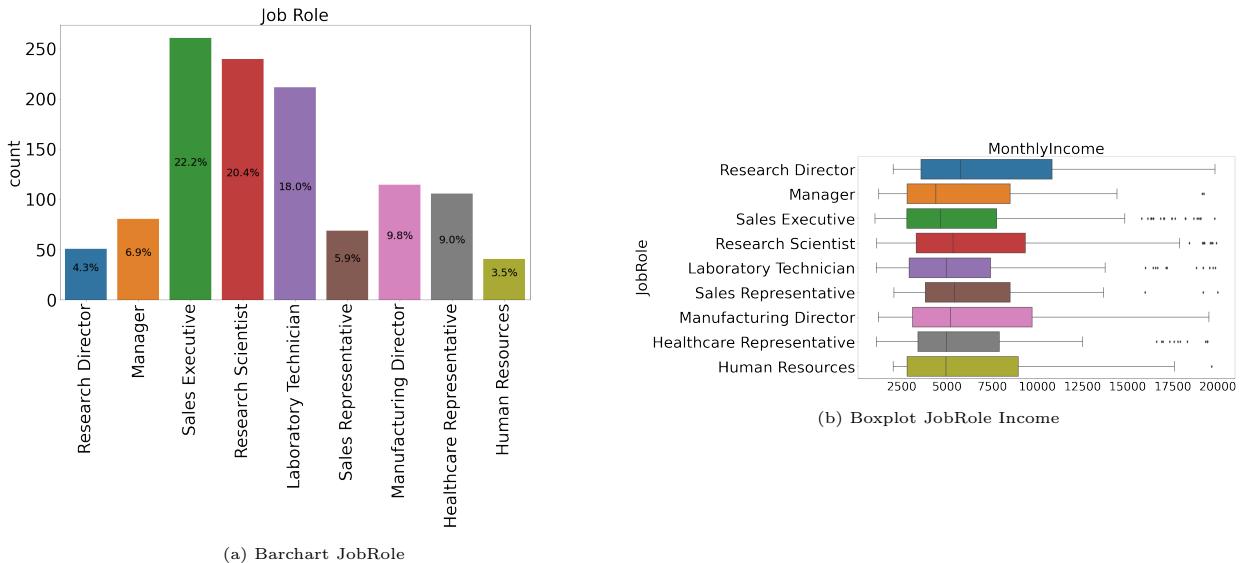


Figure 3

La maggior parte degli impiegati ha meno di 12 anni di lavoro alle spalle, con un picco di lavoratori con esperienza lavorativa tra i 10 e i 12 anni (Figura 4a). Per quanto riguarda i viaggi di lavoro, circa il 71,4% degli impiegati viaggia raramente, con una minoranza di dipendenti che non viaggia affatto o viaggia spesso (Figura 4b).

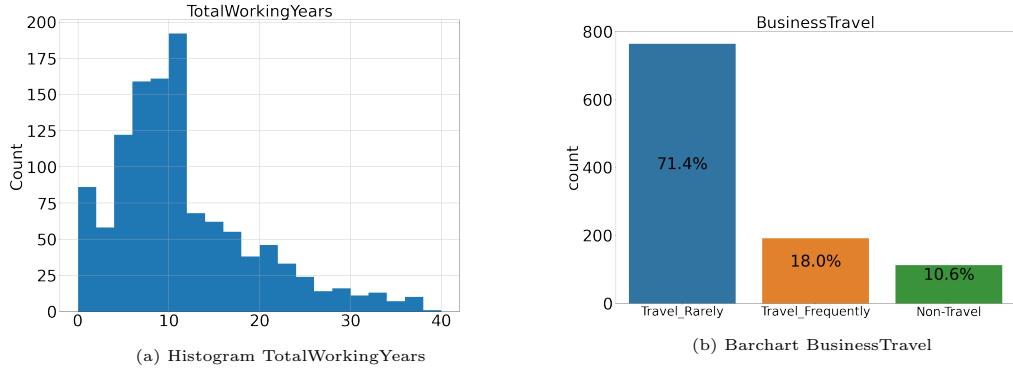


Figure 4

L'analisi del dataset in funzione dell'attrition ha evidenziato che il 16,3% del personale si dimette o si ritira dal lavoro (Figura 5a); tale percentuale corrisponde a 192 individui dei 1176 totali presenti nella collezione. Questo dato testimonia lo sbilanciamento dei dati a disposizione sull'attributo oggetto di studio. L'analisi dei successivi diagrammi ha come scopo individuare le possibili cause che comportano il fenomeno.

Analizzando i vari ruoli lavorativi in azienda (Figura 5b), è possibile osservare che il 25,5% del totale dei lavoratori nel ruolo di Laboratory Technician presentano attrition, il 24,4% di quelli in Human Resources e il 37,7% di quelli che lavorano come Sales Representative; dati che evidenziano come l'attrition sia maggiore del normale in questi 3 ruoli.

Non sembra influire sull'attrition il salario; analizzando infatti gli impiegati che guadagnano meno del 25-percentile non vi è una percentuale rilevante di attrition.

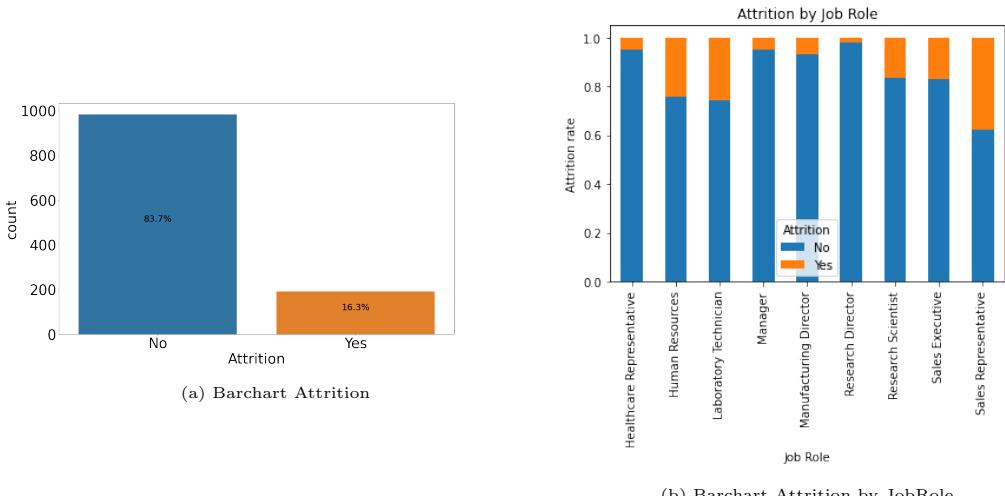


Figure 5

Dal grafico di Figura 6a, si osserva che JobLevel influisce molto sull'attrition, con alte percentuali dove questo valore è basso. Stesso discorso per quanto riguarda gli straordinari, con il 30,5% dei lavoratori che li fanno che presenta attrition, contrapposto al solo 10,6% di chi non li fa (Figura 6b). Sembra influire anche lo stato civile, con la percentuale più alta di attrition associata agli impiegati single (Figura 7a), così come il numero degli anni di lavoro (Figura 7c) e quelli nel

ruolo attuale(Figura 11b).

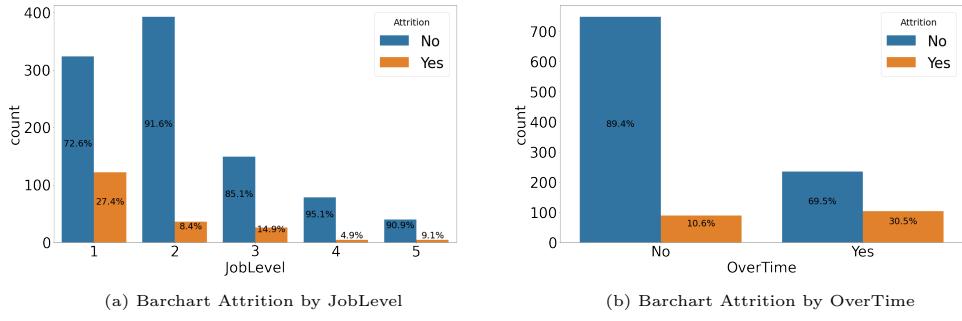


Figure 6

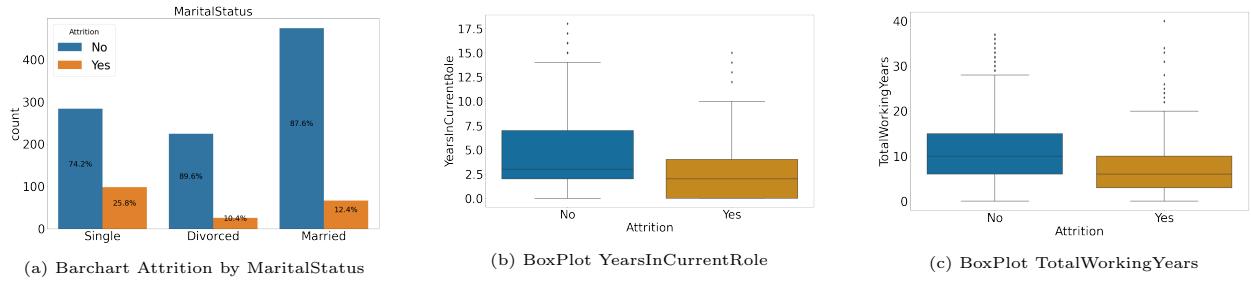


Figure 7

Di cruciale importanza sono la qualità della vita lavorativa, per la quale il 35,8% degli impiegati con indice 1 lascia il lavoro (Figura 8b), ed il coinvolgimento lavorativo dove il 32,3% degli impiegati, sempre con indice 1, abbandona (Figura 8a). Meno importanti ma comunque influenti sull'attrition sono anche gli indici che riguardano la soddisfazione del ruolo lavorativo, dell'ambiente lavorativo, e delle relazioni. In totale su 192 individui che presentano attrition, 133 hanno un basso valore dell'indice associato ai suddetti campi.

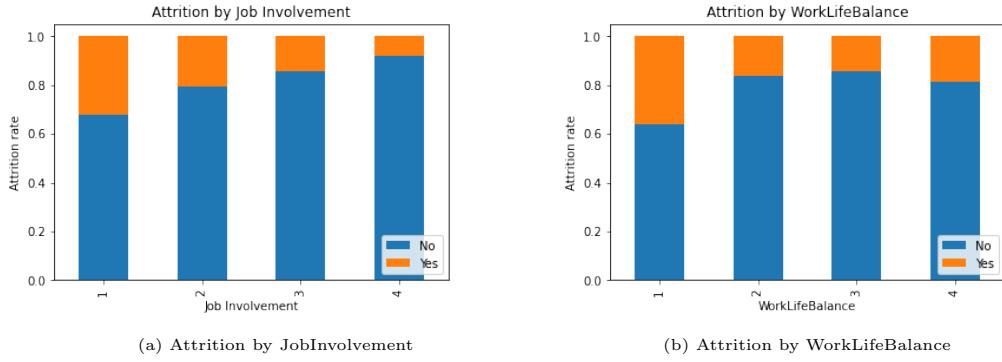


Figure 8

2.3 Data quality

In questa sezione è stata svolta un'analisi mirata ad individuare errori semantici e sintattici nei dati, missing values ed outliers. La loro presenza può infatti portare ad una brusca riduzione della qualità dei dati con importanti conseguenze negative sui risultati dello studio.

2.3.1 Errori sintattici e semantici

L'analisi effettuata andando a visualizzare i diversi valori su ogni campo, non ha rilevato la presenza di errori sintattici. Mentre dal punto di vista semantico sono state individuate delle anomalie nel campo YearsAtCompany:

- 369 record con YearsInCurrentRole > YearsAtCompany
- 169 record con YearsSinceLastPromotion > YearsAtCompany
- 321 record con YearsAtCompany > TotalWorkingYears

In particolare i 321 record in cui il campo YearsAtCompany presenta valori maggiori del campo TotalWorkingYears porta a considerare non attendibile questo campo. In aggiunta a ciò in YearsAtCompany sono presenti anche 60 missing value. E' stato quindi deciso di eliminare questo campo dallo studio della collezione.

2.3.2 Missing Values

In figura 9 sono osservabili gli attributi della collezione che presentano valori mancanti. Di seguito, per ognuno di questi campi, viene riportato il metodo deciso per rimpiazzarli.

Field	TotalMissingValue	
0	Age	176
1	BusinessTravel	107
2	Gender	59
3	MonthlyIncome	213
4	Over18	372
5	PerformanceRating	138
6	StandardHours	570
7	TrainingTimesLastYear	233
8	YearsAtCompany	60

Figure 9: Campi dove sono presenti missing value

- Age : la media di questo campo nella collezione è 37; questo valore è stato utilizzato per sostituire i valori mancanti.
- BusinessTravel : si è deciso di sostituire i valori mancanti con "Travel_Rarely" visto che il 71,5% degli individui della collezione presentano questo valore.
- Gender : avendo la collezione una percentuale di "Male" uguale al 59,4% e di "Female" pari al 40,6%, è stato deciso di inserire 35 valori "Male" e 24 valori "Female" in modo da mantenere le proporzioni.
- Over18 : attributo eliminato dalla collezione.
- StandardHours : attributo eliminato dalla collezione.
- PerformanceRating : i valori mancanti di questo campo sono stati sostituiti con il valore "3" visto che l'85% degli individui della collezione presentano questo valore.

- MonthlyIncome : per sostituire i valori mancanti di questo campo, è stata utilizzata la correlazione con Age. In particolare è stata creata una tabella (Figura 10) con dei range di età e la media dei valori di MonthlyIncome per ognuno di questi range. A seconda della categoria in cui rientra l'età dell'individuo, si può quindi inserire il valore MonthlyIncome corrispondente.
- TrainingTimesLastYear : trovando nella collezione il valore “2” con una percentuale pari al 35,9%, il valore “3” con il 34,6% e gli altri valori con meno del 9%, è stato deciso di inserire 117 valori uguali a “2” e 116 valori uguali a “3”.
- YearsAtCompany : attributo eliminato dalla collezione.

2.3.3 Outliers

Sono presenti outliers su diversi campi della collezione. In particolare sono presenti un numero consistente nei campi TotalWorkingYears, MonthlyIncome e YearsInCurrentRole, qui mostrati nei propri box-plot (Figura 11). E' stato scelto di gestirli successivamente in base al task specifico, andando a studiare ogni volta, quale fosse il metodo migliore.

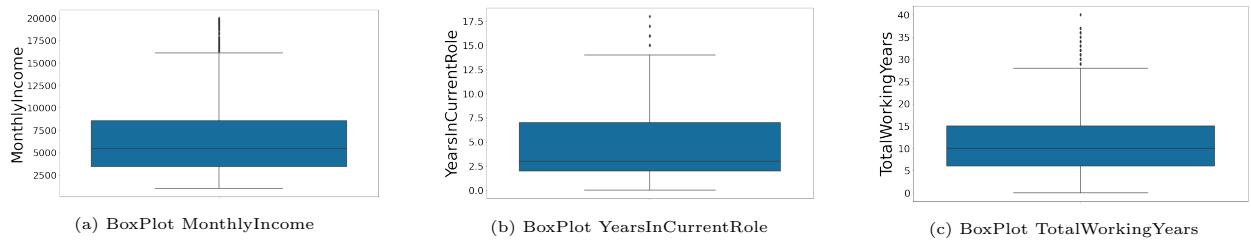


Figure 11

2.4 Correlazione ed eliminazione di variabili

Al fine di avere una visione approfondita delle relazioni tra le variabili numeriche, è stata qui calcolata la matrice di correlazione di Pearson (Figura 12). La matrice in figura è tagliata dei suoi valori sopra diagonale e presenta tutti i valori in valore assoluto.

In aggiunta in figura 13, è possibile osservare anche la matrice di correlazione per gli attributi categorici che sfrutta l'indice V di Cramer.

Osservazioni:

- correlazione di Pearson pari a 0,77 tra TotalWorkingYears e JobLevel, che ha portato alla decisione di eliminare il campo JobLevel;
- correlazione di Pearson uguale a 0,71 tra YearsInCurrentRole e YearsWithCurrManager che ha suggerito l'eliminazione dell'attributo YearsWithCurrManager;
- correlazione di Pearson pari a 0,57 tra YearsInCurrentRole e YearsSinceLastPromotion che ha condotto alla decisione di eliminare l'attributo YearsSinceLastPromotion;
- correlazione di Cramer uguale a 0,95 tra Department e JobRole che ha portato alla decisione di eliminare il campo Department.

	RangeAge	MonthlyIncome
0	(17, 30]	3839.742857
1	(30, 40]	5986.831933
2	(40, 50]	9228.730570
3	(50, 60]	10545.035714

Figure 10: MonthlyIncome medio per fasce di età

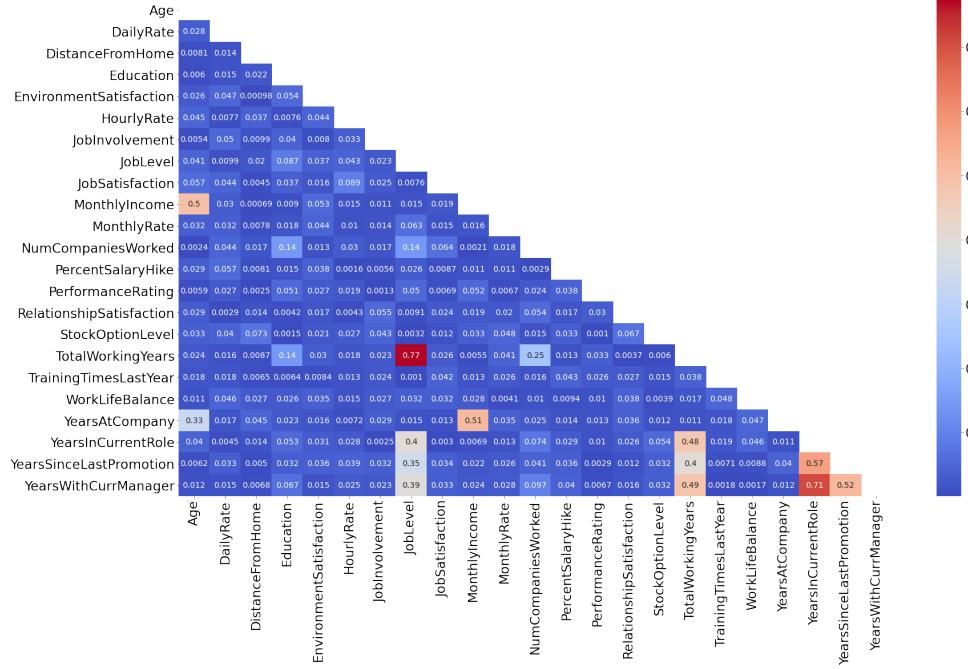


Figure 12: Matrice di correlazione di Pearson

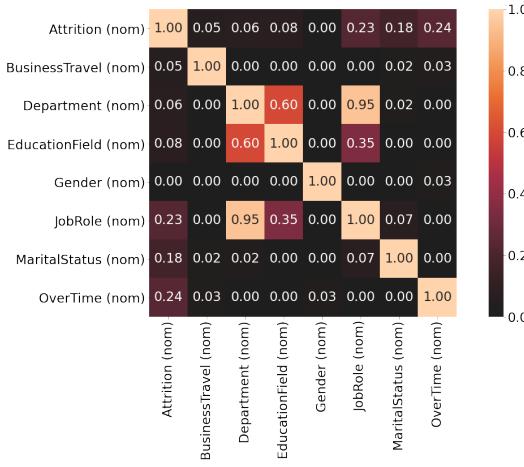


Figure 13: Matrice di correlazione di Cramer

Oltre all'eliminazione di questi campi e a quella di YearsAtCompany per i motivi già menzionati, è stato deciso di eliminare anche StandardHours, Over18, HourlyRate, DailyRate e MonthlyRate. StandardHours, Over18 sono stati tolti dall'analisi perché presentavano sempre e solo lo stesso valore su tutti gli individui della collezione. HourlyRate, DailyRate e MonthlyRate invece sono stati invece eliminati dopo un attento studio: questi attributi sono riferiti al costo orario, giornaliero e mensile che l'azienda spende per un dipendente, includendo quindi vari costi oltre allo stipendio netto. Riflettendo sul significato di questi valori, è stato ritenuto che non possano avere un'influenza sull'attrition, essendo costi aziendali che non riguardano il dipendente (la cui paga netta è espressa dall'attributo MonthlyIncome). In più, sono state riscontrate delle difficoltà nell'interpretare i dati presenti nei 3 attributi, non presentando essi nessuna correlazione tra di loro.

2.5 Trasformazione delle variabili

Almeno inizialmente non sono state eseguite trasformazioni dei dati categorici in numerici discreti, normalizzazioni o discretizzazioni di attributi numerici continui. Questo perchè è sembrato più

giusto valutare successivamente, a seconda del task oggetto di studio, quali trasformazioni effettuare per rendere migliore l’analisi corrente.

Per ridurre la dimensionalità della collezione, sono stati però uniti gli attributi JobInvolvement, JobSatisfaction, EnvironmentSatisfaction, RelationshipSatisfaction e WorkLifeBalance. Questi 5 attributi riguardano tutti lo stato emotivo che il dipendente prova verso il proprio lavoro e presentano tutti valori che vanno da 1 a 5. Si è deciso quindi di sommare i valori di questi 5 attributi e creare un nuovo attributo chiamato Satisfaction.

3 Clustering

3.1 Scelta degli attributi

In questa sezione del progetto, si è provato a raggruppare i dati di alcuni attributi per vedere se ci fossero delle strutture interne, magari in grado di classificare gli individui in base all’attrition. Sono stati quindi esaminate 3 tecniche di clustering diverse su 3 attributi scelti dalla collezione. Gli attributi scelti sono stati: Satisfaction, TotalWorkingYears e YearsInCurrentRole. La scelta di questi 3 attributi è maturata dopo aver visto i loro grafici box-plot tenendo conto dell’attrition (Figura 14). Si può notare infatti da questi 3 grafici come i valori del primo, del secondo e del terzo quartile siano inferiori negli individui con attrition, dato che porta a ritenere una possibile correlazione meritevole di studio.

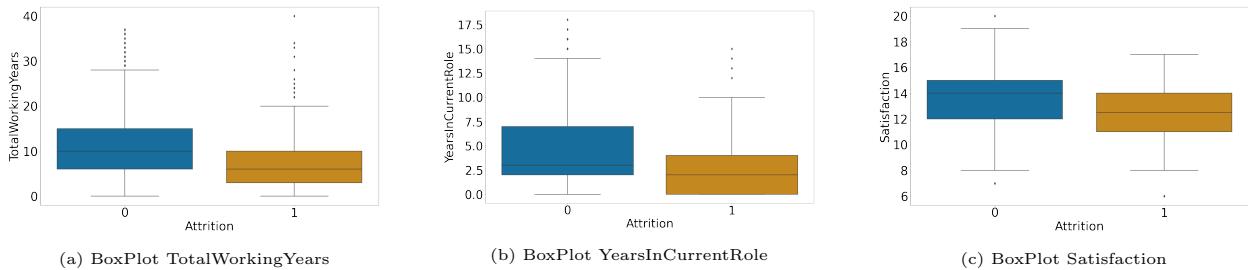


Figure 14

Al fine di una migliore analisi, sono stati eliminati gli outliers di questi 3 attributi, portando il numero di record da 1176 a 1013. Questa eliminazione non comporta variazioni significative nella percentuale di individui con attrition nella collezione, che passa da 16,3% a 17,5%. I dati prima dell’analisi sono stati anche normalizzati utilizzando la tecnica del “RobustScaler” che, non basandosi sulla media, si adatta meglio con le distribuzioni diverse da quelle di Gauss (“YearsInCurrentRole” ha una distribuzione bimodale). Per la metrica infine, è stato deciso di valutare per ogni diversa tecnica di clustering quale fosse la migliore tra la distanza Euclidea e quella Manhattan.

3.2 K-Means

3.2.1 Identificazione del migliore valore di K

Nello studio del clustering attraverso l’algoritmo del K-Means, risulta di fondamentale importanza la scelta di 2 parametri: il numero dei cluster e la posizione iniziale dei centroidi. Per individuare il numero di cluster K, è stato utilizzato il knee method. In figura 15a, sono stati quindi riportati su un grafico gli andamenti dell’SSE e della Silhouette al variare di K.

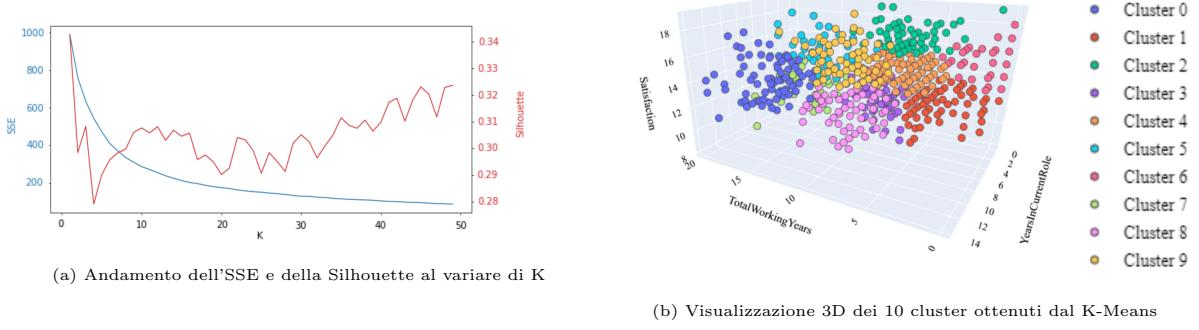


Figure 15

Analizzando il grafico è stato scelto un numero di cluster $K=10$; questo valore si trova infatti in prossimità del punto di gomito e presenta inoltre un buon valore della Silhouette (SSE=310 e Silhouette=0.30)

Una volta scelto il K si è potuto eseguire l'algoritmo, selezionando 50 configurazioni iniziali random dei centroidi in modo che l'algoritmo trovasse in automatico la migliore. Questa operazione, eseguita con la distanza Euclidea, ha generato la configurazione di cluster visibile in figura 15b.

3.2.2 Caratterizzazione dei cluster ottenuti

Al fine di studiare a fondo i 10 cluster ottenuti e porteli confrontare in base all'attrition, sono stati realizzati 2 diagrammi: un primo dove si possono osservare le coordinate dei centroidi per ognuno dei 3 attributi (Figura 16a) ed un secondo in cui è riportata per ogni cluster la somma normalizzata degli individui con o senza attrition (Figura 16b).

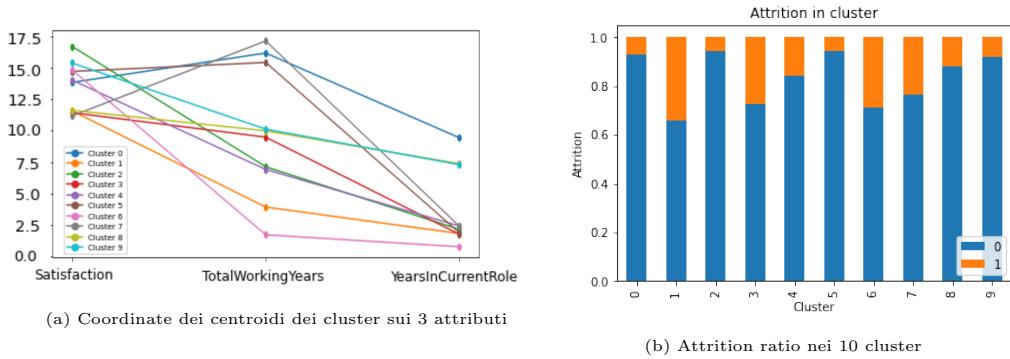


Figure 16

Si può notare che sull'attributo Satisfaction si viene a creare una sorta di scissione tra cluster che presentano un valore più basso di soddisfazione e altri che presentano un valore medio-alto. Andando ad analizzare i cluster con valori più bassi (1-3-7-8) si riscontra una maggiore percentuale relativa di attrition, ad eccezione del cluster numero 8. Spostandoci sull'attributo TotalWorkingYears, si nota che si vengono a creare 3 diversi gruppi di cluster. L'influenza di questo attributo sull'attrition, non è di facile lettura; risultano infatti fortemente influenzati dall'attrition sia il cluster 7 che il cluster 6, le cui coordinate dei centroidi sull'attributo corrispondono al valore più alto e più basso. Un'influenza maggiore sembra invece causata dall'attributo YearsInCurrentRole, dove i cluster possono essere suddivisi in 2 gruppi: un primo gruppo composto dai cluster 0-8-9, con alti valori dei centroidi sull'attributo, caratterizzato da bassa percentuale di attrition, ed un secondo gruppo, con valore dei centroidi basso, dove vi sono sia cluster con alte percentuali di attrition, sia cluster con basse. Analizzando i cluster all'interno di questo gruppo, è possibile notare che i cluster 2,4,5 caratterizzati da basse percentuali di attrition, hanno alti valori su Satisfaction. Al contrario i cluster 1,3,7, in cui i valori su Satisfaction sono bassi, hanno alte percentuali di attrition. Par-

ticolare è il comportamento del cluster 6 che possiede il più basso valore su YearsInCurrentRole: sebbene presenti un valore di Satisfaction alto, è uno dei cluster che presenta maggior attrition.

In conclusione, dai 10 cluster ottenuti non è stato ottenuto nessun cluster con una percentuale schiacciante di impiegati che presentino attrition, ma è stata comunque notata una grande influenza da parte degli attributi Satisfaction e YearsInCurrentRole. Bassi valori in entrambi questi attributi, conducono ad una più alta percentuale relativa di attrition con picchi che superano il 30%. Al contrario, alti valori in questi 2 attributi, riducono l'attrition sotto il 10%.

3.3 DBScan

3.3.1 Studio dei parametri per il clustering

Per eseguire il clustering con la tecnica del DBScan, è necessario scegliere attentamente due parametri: MinPoints e Epsilon. La ricerca di questi 2 parametri è iniziata andando ad individuare un range in cui selezionare l'Epsilon migliore con il metodo del K-Nearest neighbor, utilizzando sia la distanza Manhattan che la distanza Euclidea.

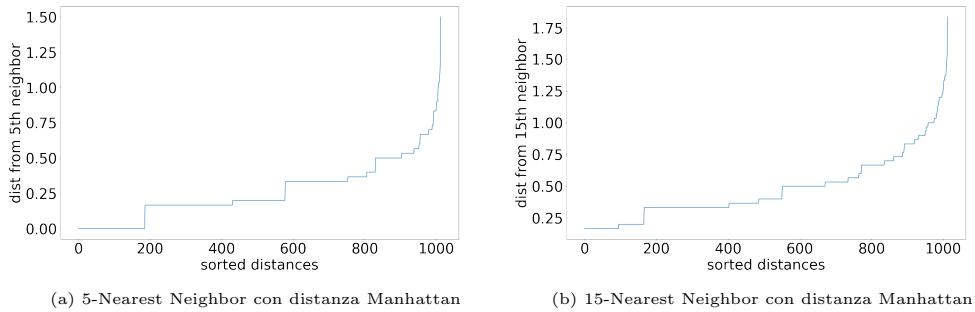


Figure 17

Il metodo è stato provato con diversi valori di K, cioè calcolando la distanza del k-esimo “vicino” di ogni punto del dataset. Dai due grafici in figura 17, appare evidente che molti individui presentano gli stessi valori su tutti e 3 gli attributi; questo è intuibile dalla curva “a scalini” della figura 17a: il valore zero della curva su quasi 200 punti, sta ad indicare che circa 200 individui della collezione potrebbero avere fino 5 individui con gli stessi valori. Dalla figura 17b, si vede invece che aumentando il valore di K, diminuisco gli scalini sulla curva, ma restano comunque presenti. Non si riscontrano invece sostanziali differenze tra le due differenti misure provate. Si è quindi scelto di utilizzare la distanza Manhattan, per il resto dell’analisi.

Successivamente si è provato a selezionare il miglior compromesso tra MinPoints e Epsilon, in base al valore della Silhouette (Figura 18). I valori di Epsilon qui provati vanno da 0.4 a 0.8 e mostrano la variazione della Silhouette al variare di K. Dalla figura è possibile osservare che il valore della Silhouette più alto si ottiene con Epsilon=0.8 e MinPoints uguale ad un valore di 3 o 4.

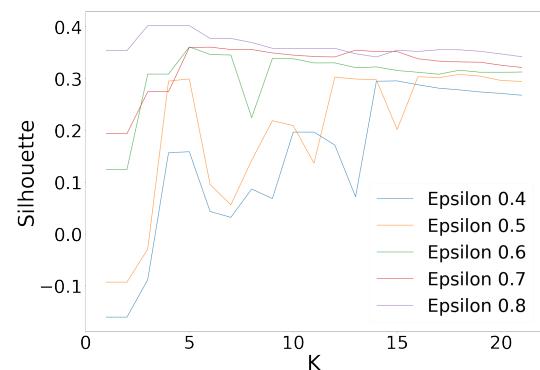


Figure 18: Variazione del valore della Silhouette rispetto al Min-Points e all’Epsilon

3.3.2 Caratterizzazione e interpretazione dei cluster ottenuti

Andando ad eseguire il DBScan con i parametri individuati precedentemente, viene restituito un unico grande cluster con 1011 punti e 2 noise points. Provando a cambiare i parametri utilizzati,

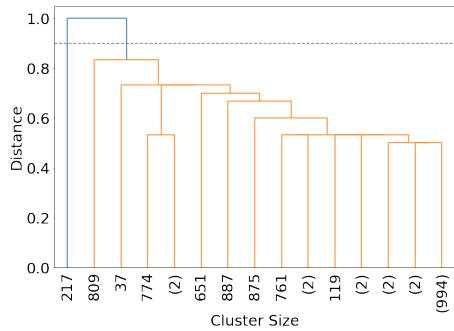
con Epsilon=0.7 e MinPoints=13 si forma ancora un unico grande cluster. Un'attenta analisi sulle motivazioni di questi risultati, ha portato a concludere che, data la natura discreta degli attributi utilizzati per il clustering, il range abbastanza ristretto su cui si distribuiscono i punti ed i molti punti sovrapposti, l'analisi del clustering attraverso il metodo del DBScan non sia la più corretta. Essa infatti, basandosi sulla densità dei punti, porterà facilmente alla creazione di un unico cluster con questi attributi, impedendo l'analisi mirata ad individuare un possibile schema discriminatorio per l'attrition.

3.4 Hierarchical

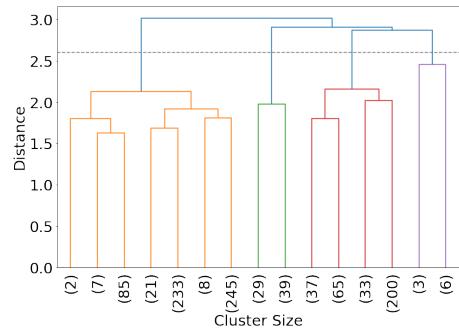
3.4.1 Visualizzazione e discussione di dendrogrammi differenti ottenuti con algoritmi differenti

Il clustering gerarchico sul dataset è stato effettuato utilizzando i metodi Complete Linkage, Single Link e Group Average. Con tutti e 3 i metodi, sono state provate sia la distanza Manhattan che la distanza Euclidea, ma variando di veramente poco i risultati, si è scelto la distanza Manhattan per svolgere l'analisi.

In figura 19a è rappresentato il dendrogramma relativo al Single-linkage che come si può notare produce molti cluster contenenti un unico record. Il dendrogramma con il metodo dell'Average-linkage è invece riportato in figura 19b: utilizzando come threshold il valore 2.6 si ottengono 4 cluster, di cui 2 contengono relativamente pochi records (9 e 68), e gli altri due sono poco rilevanti in termini di informazione sull'attrition.



(a) Dendrogramma Single-linkage



(b) Dendrogramma Average-linkage

Figure 19

Il risultato migliore si è ottenuto invece con il metodo Complete-linkage (figura 20). Su quest'ultimo è stato effettuato un taglio a distanza 4.3 dal quale si ottengono 7 cluster (figura 21b). Il numero di cluster scelto deriva da un'analisi della silhouette: con 7 cluster si ottiene un valore di 0.22, minore solo del valore di 0.32 ottenuto con 2 cluster che però sono stati valutati poco significativi per lo studio dell'attrition.

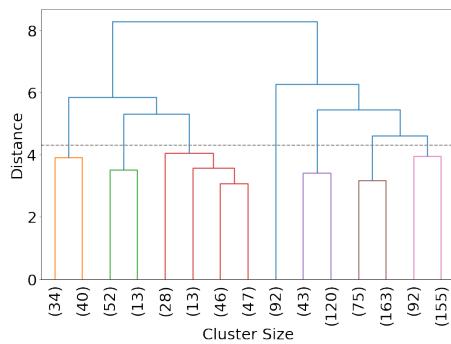


Figure 20: Dendrogramma Complete-linkage

Analizzando i 7 cluster ottenuti si può notare che la percentuale più alta di attrition si riscontra nei cluster 4 (30.1%) e 5 (45.7%) che rappresentano quella fetta di lavoratori con bassi valori in

TotalWorkingYears e YearsInCurrentRole. Il cluster 5 presenta inoltre bassi valori anche di Satisfaction. La percentuale diminuisce invece significativamente nel cluster 0 (7.5%), che è caratterizzato da valori alti di TotalWorkingYears e YearsInCurrentRole, e nel cluster 1 (8.5%) contenente records con alta Satisfaction.

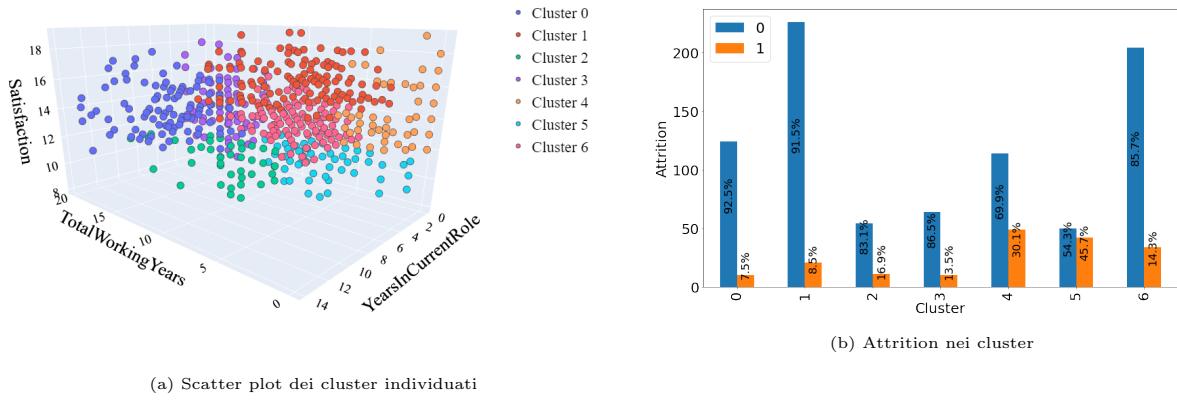


Figure 21

3.5 Valutazione finale del miglior approccio di clustering

Dall'analisi del dataset con le 3 diverse tecniche di clustering non è stato possibile individuare una chiara struttura interna dei dati in grado di classificare in modo chiaro il fenomeno dell'attrition in azienda.

L'analisi con la tecnica del DBScan è risultata del tutto inadatta con i dati a disposizione, andando ad individuare 1 solo cluster, a meno di non ridurre considerevolmente la Silhouette.

Le tecniche del K-means e del clustering gerarchico (metodo Complete) hanno invece prodotto rispettivamente 10 e 7 cluster. Il clustering attraverso il K-means presenta un valore della Silhouette pari a 0,30 contro lo 0,22 del clustering gerarchico; un valore più alto della Silhouette testimonia una maggiore coesione e separazione tra i cluster e quindi una migliore distinzione di gruppi di individui. Per questi motivi la tecnica di clustering del K-means è da preferirsi in questa analisi.

4 Classification

4.1 Decision Tree

4.1.1 Introduzione

In questa sezione verrà descritto in che modo è stato svolto il task di Classificazione con il metodo del Decision Tree. Il dataset utilizzato per lo studio dell'Attrition risulta diviso in due sottoinsiemi: il training set, che è stato utilizzato per allenare il modello, ed il test set, utilizzato invece per valutare le performance del modello scelto.

Il training set contiene 1176 record di cui solo 192 presentano Attrition, si tratta quindi di un insieme di dati piuttosto sbilanciato. Per sopprimere a questa problematica, che non permette di effettuare un buon training del modello, sono state adottate diverse strategie:

- Oversampling e undersampling, tecniche che permettono di bilanciare i dati, inserendo record finti, derivanti dagli originali, per la classe marginale, oppure togliendo record della classe dominante.
- Validazione del modello con la metrica ‘balanced accuracy’, efficace con i dataset sbilanciati in quanto combina Sensitivity (recall) e Specificity (true negative rate).
- Assegnando dei pesi alle classi attraverso il parametro `class_weight` del `DecisionTreeClassifier`.

criterion	[gini,entropy]
max_depth	[2,15]
min_samples_split	[5,50]
min_samples_leaf	[5,50]

Table 2: Griglia iperparametri

Per la ricerca degli iperparametri è stata utilizzata la Grid-Search con la quale viene effettuata una Stratified K-fold cross validation con 15 subset per ottenere una validazione più affidabile e allo stesso tempo mantenere le proporzioni dell'attrition nei subset. Gli iperparametri e i rispettivi range provati sono riportati in tabella 2.

Di seguito andremo ad analizzare 3 modelli con caratteristiche differenti per poi delineare il migliore e valutarne le performance sul test set.

4.1.2 Modello 1

Col primo modello non sono state messe in atto strategie per lo sbilanciamento del dataset, infatti i risultati ottenuti sono peggiori rispetto agli altri modelli ottenuti.

In tabella 3 sono mostrati i parametri utilizzati risultanti dalla GridSearch. Sono stati inoltre riportati in tabella 4 i diversi score ottenuti sul Training Set e con la K-fold cross validation, sui quali verranno confrontati i diversi modelli per scegliere il migliore.

Salta subito all'occhio che, data la dominanza dei record che non presentano Attrition, il modello costruito tende ad assegnare i record alla classe 'No'. Infatti la misura Recall risulta piuttosto bassa in quanto più della metà dei record che presentano Attrition vengono predetti in modo errato. Inoltre con la balanced accuracy si ottiene un valore più significativo rispetto all'accuracy standard che avrebbe mostrato uno score molto più alto per l'alto numero di 'No' predetti correttamente.

criterion	entropy
max_depth	8
min_samples_split	15
min_samples_leaf	10

Table 3: Iperparametri scelti

	Bal Accuracy	F1	Recall	Precision
Training	0.72	0.58	0.46	0.76
Validation	0.65	0.42	0.37	0.54

Table 4: Score modello

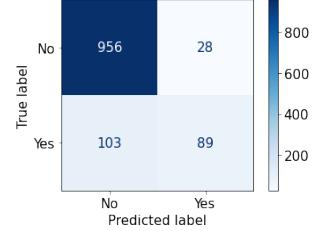


Figure 22: Confusion Matrix sul train set

4.1.3 Modello 2

Il secondo modello è stato ottenuto dal training su un set di dati bilanciato. Per far questo è stato prima diviso il dataset in Training set e Validation set; a questo punto sono state utilizzate le tecniche di oversampling e undersampling solo sul Training set ottenendo 344 record che presentano Attrition e 344 no. E' stato invece mantenuto lo sbilanciamento nel Validation Set in modo da poter valutare i modelli con diversi iperparametri su un dataset reale.

Il risultato sul Training set è fortemente influenzato dai record fittizi e perciò non permette di ottenere informazioni particolari, mentre dallo score sul Validation set si nota una migliore Recall dovuta al training del modello con più record della classe 'Yes' che permettono di predire meglio i record che presentano Attrition.

criterion	gini
max_depth	5
min_samples_split	20
min_samples_leaf	5

Table 5: Iperparametri scelti

	Bal Accuracy	F1	Recall	Precision
Training	0.83	0.81	0.76	0.87
Validation	0.69	0.45	0.53	0.39

Table 6: Score modello

4.1.4 Modello 3

In questo caso non sono state applicate le tecniche dell'over/undersampling come in precedenza, ma si è utilizzato invece il parametro `class_weight` del `DecisionTreeClassifier` settato a `balanced`, assegnando i pesi alle classi in base alla loro frequenza.

Il modello ottenuto presenta il miglior score su Recall riuscendo a predire correttamente più di 2/3 dei record con etichetta 'Yes' del Validation set, oltre ad avere una buona accuracy con score 0.74.

criterion	gini
max_depth	3
min_samples_split	5
min_samples_leaf	10
class_weight	balanced

Table 7: Iperparametri scelti

	Bal Accuracy	F1	Recall	Precision
Training	0.75	0.53	0.68	0.43
Validation	0.74	0.51	0.67	0.41

Table 8: Score modello

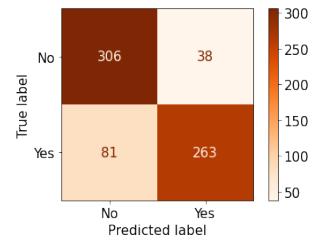


Figure 23: Confusion Matrix sul train set

4.1.5 Scelta e valutazione sul test set del miglior modello

	Bal Accuracy	F1	Recall	Precision
Modello 1	0.65	0.42	0.37	0.54
Modello 2	0.69	0.45	0.53	0.39
Modello 3	0.74	0.51	0.67	0.41

Table 9: Score sul Validation set dei modelli

Dal confronto tra i diversi modelli, il modello 3 è risultato il migliore in quanto ha prodotto gli score più alti per la maggior parte delle metriche, secondo solo al modello 1 in termini di Precision a causa del fatto che quest'ultimo predice con meno frequenza la classe 'Yes'.

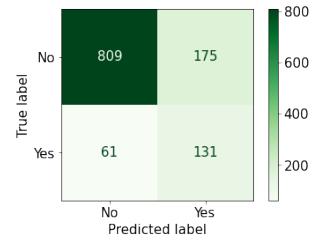


Figure 24: Confusion Matrix sul train set

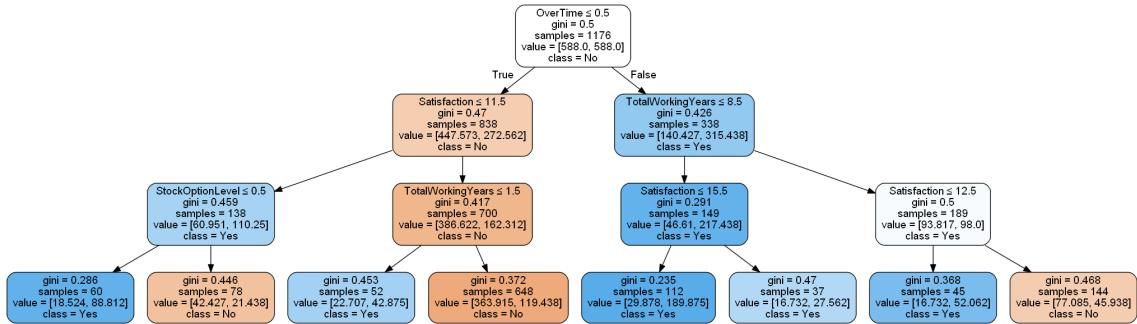


Figure 25: Decision Tree modello 3

Il modello scelto è caratterizzato da una profondità massima di 3 (che permette di mostrare il Decision Tree costruito nella sua interezza) indice di una complessità ridotta, ciò riduce il rischio di overfitting.

Nella costruzione del Decision Tree sono stati presi in considerazione solo 4 attributi (di cui è mostrata l'importanza nella tabella 12), mentre gli altri non contribuiscono ad incrementare l'information gain.

Si è applicato il modello al test set per ottenere una stima sulle performance, riscontrando un valore di balanced accuracy di 0.72. Dalla confusion matrix sul test set si nota che i record della classe ‘No’ predetti correttamente sono 214 su 249, mentre per la classe ‘Yes’ sono 26 su 45 (quasi il 60%).

Attributo	Importanza
OverTime	0.334561
StockOptionLevel	0.118095
TotalWorkingYears	0.248877
Satisfaction	0.298466
Altri	0.000000

Table 10: Importanza attributi

	Bal Accuracy	F1	Recall	Precision
Train set	0.75	0.53	0.68	0.43
Test set	0.72	0.49	0.58	0.43

Table 11: Score su train e test set

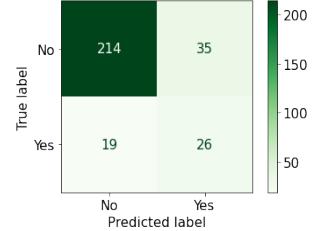


Figure 26: Confusion Matrix sul test set

4.2 KNN

4.2.1 Introduzione

La classificazione dei dati mediante l'algoritmo KNN ha dovuto affrontare gli stessi problemi di sbilanciamento del label dell'Attrition, già osservati durante l'analisi del Decision Tree. Un primo metodo, per cercare di attenuare questa situazione, è stato quello di classificare i nuovi record non più semplicemente in base alla classe dominante tra i K “vicini” scelti, ma dare anche dei pesi in base alla distanza dal valore del record da classificare. Questa tecnica permette un'analisi più accurata dei dati ma da sola difficilmente riesce a compensare un forte sbilanciamento dei dati. Per questo è stato deciso di utilizzare anche le tecniche di Oversampling e Undersampling, in modo da partire già con un set di dati bilanciati. Nell'algoritmo del KNN, un altro problema da non sottovalutare è il “Curse of Dimensionality”: si possono infatti ottenere risultati privi di significato se vengono analizzati contemporaneamente un gran numero di attributi. Per questo motivo abbiamo eseguito analisi diverse su due set di attributi che dalle precedenti analisi sembravano maggiormente influenzare l'attrition:

- Gruppo 1: PercentSalaryHike, StockOptionLevel, TotalWorkingYears, YearsInCurrentRole, NumCompaniesWorked, Satisfaction.

- Gruppo 2: TotalWorkingYears, YearsInCurrentRole, Satisfaction.

Infine, prima di iniziare le analisi, è stato necessario normalizzare i dati al fine di dare la stessa importanza a tutti gli attributi utilizzati. Per fare questo è stata utilizzata la tecnica del RobustScaler, che si basa sui percentili e non viene quindi influenzata da outliers e si adatta meglio a distribuzioni diverse da quella di Gauss.

4.2.2 Selezione dei parametri

Durante le analisi sono stati quindi ricercati i migliori parametri per far rendere al massimo le prestazioni di questo algoritmo; oltre a questo è stato anche necessario studiare su quali attributi l'algoritmo desse i migliori risultati e quale fosse il miglior training set di partenza tra quello originale e quello con i dati modificati dalle tecniche di Oversampling e Undersampling. Per questi motivi sono state condotte le ricerche dei migliori parametri su 4 combinazioni diverse di attributi e training set:

- Combo 1: attributi del gruppo 1 con dati del training set originale. I migliori parametri sono stati ricercati utilizzando la tecnica della cross validation.
- Combo 2: attributi del gruppo 2 con dati del training set originale. I migliori parametri sono stati ricercati utilizzando anche qui la tecnica della cross validation.
- Combo 3: attributi del gruppo 1 con dati modificati dalle tecniche di Oversampling e Undersampling. Per una corretta valutazione, il training set originale è stato suddiviso in 2 (mantenendo in entrambi la stessa percentuale di attrition), con il 70% dei dati che è stato sottoposte alle modifiche al fine di aver un bilanciamento dell'attrition ed un 30% dei dati che invece non ha subito modifiche e ha rappresentato il validation set. Questo ha permesso una valutazione della strategia di analisi non influenzata dalle modifiche di Oversampling e di Undersampling.
- Combo 4: attributi del gruppo 2 con dati modificati dalle tecniche di Oversampling e Undersampling con le stesse divisioni viste nella combinazione precedente per la corretta valutazione.

Per queste 4 combinazioni sono stati ricercati i migliori parametri di “K-neighbors” (il numero di vicini da considerare), del “weight”, uniform o distance (cioè se applicare o no un peso in base alla distanza per la classificazione del record) e della “measure” (cioè se utilizzare come misura la distanza Euclidea o quella Manhattan). I migliori parametri per ognuna delle combinazioni sono riportati in tabella 12; sono inoltre riportati i risultati dell'applicazione del validation set per le combinazioni 3 e 4, mentre per le combinazioni 1 e 2 sono riportati i risultati medi ottenuti grazie alla cross validation.

	K-Neighbors	Weight	Measure	Bal Accuracy	F1	Recall	Precision
Combo 1	3	Uniform	Manhattan	0.61	0.35	0.28	0.49
Combo 2	3	Distance	Euclidean	0.59	0.30	0.25	0.38
Combo 3	10	Uniform	Euclidean	0.67	0.42	0.59	0.33
Combo 4	17	Distance	Euclidean	0.65	0.39	0.55	0.30

Table 12: Iperparametri scelti e score

Dalla lettura della tabella 12, appare evidente che la miglior combinazione su cui eseguire l'algoritmo del KNN è la numero 3, cioè quella dove si è applicato Oversampling e Undersampling sui dati degli attributi del gruppo 1. Per questa combinazione i migliori parametri sono K=10, applicare un peso uniforme ed utilizzare la metrica Euclidean.

4.2.3 Valutazione sul test set

Al fine della valutazione finale, è stato testato l'algoritmo con la combinazione 3 (e i suoi migliori parametri) sul test set. I risultati sono riportati in tabella 13, mentre in figura 27 è riportata la confusion matrix.

	Bal Accuracy	F1	Recall	Precision
Test set	0.61	0.33	0.42	0.27

Table 13: Score test set

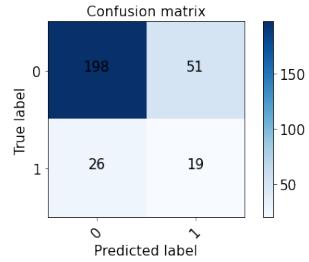


Figure 27: Confusion Matrix sul test set

4.3 Discussione sul miglior modello

Dai risultati ottenuti, è possibile stabilire che il modello 3 del Decision tree è quello che permette una migliore classificazione della classe dell'attrition sui nuovi record. Questo modello presenta infatti il valore migliore di balanced accuracy sul validation set, testimonianza della migliore capacità relativa di classificare i nuovi record caratterizzati o meno dall'attrition. Facendo un'analisi più profonda del problema, risulta però che anche questo modello non sia in senso assoluto un buon modello: lo studio infatti dovrebbe mirare ad individuare i lavoratori che sono a rischio di attrition, in modo che l'azienda possa gestire preventivamente la situazione. Appare quindi più giusto andare ad analizzare i valori della Recall per quanto riguarda i record con attrition si, in modo da capire quanti di essi la classificazione non è riuscita ad individuare. Il modello 3 del Decision Tree presenta il miglior valore anche in questa misura, ma il valore di 0,67 non è un valore di grande affidabilità. In più il valore scende ulteriormente sul test set arrivando a 0,58, di poco superiore ad una assegnazione casuale dell'attrition sui nuovi record caratterizzati dal fenomeno. Per questi motivi è stato concluso che gli algoritmi di Decision Tree e KNN non sembrano in grado di fornire un grande aiuto allo studio del fenomeno.

5 Association Rules Mining

5.1 Data preparation

I dati utilizzati per questo task sono stati sottoposti a tutte le fasi di "pulizia" viste nella sezione di Data Understanding, eccetto per quanto riguarda la fase di riempimento dei missing value: l'association rules mining offre infatti la possibilità di andare ad individuare ed a riempire i valori mancanti nel dataset. Sarà quindi valutata la possibilità di andarli a riempire più accuratamente con questa analisi. Il task dell'association rules mining inoltre, lavora bene con attributi che presentano un numero ristretto di valori possibili; per questo motivo i valori degli attributi numerici continui o discreti (dove presentavano un numero consistente di possibili valori) sono stati raggruppati in intervalli come mostrato nelle tabelle 14 e 15. Gli intervalli sono stati individuati utilizzando la tecnica di Clustering gerarchico con il metodo "Complete Linkage", cercando di dividere i valori di ogni attributo in 4-5 cluster.

PercentSalaryHike	NumCompaniesWorked	YearsInCurrentRole	TotalWorkingYears
[11; 13]	[0; 2]	[0; 2]	[0; 10]
[14; 18]	[3; 5]	[3; 6]	[11; 18]
[19; 20]	[6; 7]	[7; 10]	[19; 31]
[21; 25]	[8; 9]	[11; 13]	[19; 31]

Table 14: Intervalli degli attributi numerici modificati

Satisfaction	DistanceFromHome	Age	MonthlyIncome
[6; 9]	[1; 6]	[18; 24]	[1009; 6244]
[10; 11]	[7; 11]	[25; 32]	[6272; 9419]
[12; 14]	[12; 18]	[33; 43]	[9547; 14411]
[15; 16]	[19; 22]	[44; 51]	[14814; 19999]
[17; 20]	[23; 29]	[52; 60]	

Table 15: Intervalli degli attributi numerici modificati

5.2 Frequent Patterns

La prima parte del task dell’association rules mining consiste nell’individuare i Frequent Itemsets; è stato utilizzato per questo scopo l’algoritmo Apriori con numero minimo di item per itemset pari a 3 ed un valore di threshold del Support pari al 10%. Il risultato ha prodotto ben 2812 Frequent Items, con quelli con valore di Support maggiore che sono riportati in tabella 16. Per cercare di rappresentare questi risultati in una forma più compatta, sono stati ricercati anche i Closed Frequent Itemset ed i Maximal Frequent Itemset. L’analisi, con gli stessi parametri applicati all’algoritmo, ha prodotto 2740 Closed Frequent Itemset e 1277 Maximal Frequent Itemset. I closed Frequent Itemset sono riportati in tabella 16 insieme ai Frequent Itemset, non discostandosi da essi negli itemset con maggior Support, mentre i Maximal Frequent Itemset sono riportati in tabella 17.

Support	(Closed) Frequent Itemset
47.9%	OverTime=No, PerformanceRating=3, Attrition=No
41.8%	BusinessTravel=Travel_Rarely, OverTime=No, Attrition=No
41.2%	BusinessTravel=Travel_Rarely, PerformanceRating=3, Attrition=No
40.3%	TotalWorkingYears=[0; 10], OverTime=No, Attrition=No
38.1%	TotalWorkingYears=[0; 10], PerformanceRating=3, Attrition=No

Table 16: Closed Frequent Itemset e Frequent Itemset con Support più elevato

Support	Maximal Frequent Itemset
15.6%	OverTime=No, PerformanceRating=3, Attrition=No, BusinessTravel=Travel_Rarely, Satisfaction=[12; 14]
15.4%	OverTime=No, PerformanceRating=3, Attrition=No, BusinessTravel=Travel_Rarely, Marital_Status=Married
14.9%	OverTime=No, PerformanceRating=3, Attrition=No, StockOptionLevel=0, Marital_Status=Single
14.6%	OverTime=Si, PerformanceRating=3, Attrition=No
14.6%	OverTime=No, PerformanceRating=3, Attrition=No, StockOptionLevel=1, Marital_Status=Married

Table 17: Maximal Frequent Itemset con Support più elevato

I due Frequent Items più diffusi presentano entrambi i valori OverTime=No e Attrition=No, mentre cambiano il terzo valore che nel primo è PerformanceRating=3 e nel secondo è BusinessTravel=Travel Rarely. Tutti e 4 questi valori si ritrovano anche nei 2 Maximal Frequent Items più diffusi. Questi risultati sono facilmente interpretabili dalle elevate percentuali di distribuzione di questi valori nel dataset, già viste nella sezione di Data Understanding.

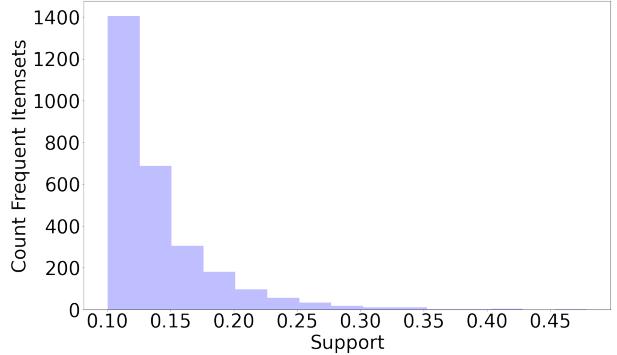


Figure 28: Istogramma Frequent Itemset

In figura 28 è riportato un istogramma che aiuta a capire meglio la distribuzione dei 2812 Frequent Itemset individuati in base alla soglia di Support. Come è logico aspettarsi il numero di Frequent Itemset decresce rapidamente all’aumentare del valore di Support con ”sol” 237 itemset maggiori di Support=20% e 29 maggiori di Support=30%.

5.3 Association Rules

Per l'estrazione di Association Rules è stata fissata la threshold min_sup=10 e il parametro zmin=3, che indica la cardinalità minima degli itemsets frequenti presi in considerazione.

Si è partiti utilizzando una confidence minima del 60% per analizzare in che modo variasse il numero di regole. In figura 29a si può notare che impostando min_conf=60 si ottengono poco meno di 14.000 regole, numero che cala progressivamente fino ad arrivare a 91 per le association rules con confidence 100%. Di fianco è riportato l'istogramma che rappresenta la distribuzione di regole in funzione della confidence, di cui la maggior parte sono concentrate sotto l'80%. Per l'estrazione di association rules prenderemo in considerazione quelle superiori a questa soglia in quanto più significative.

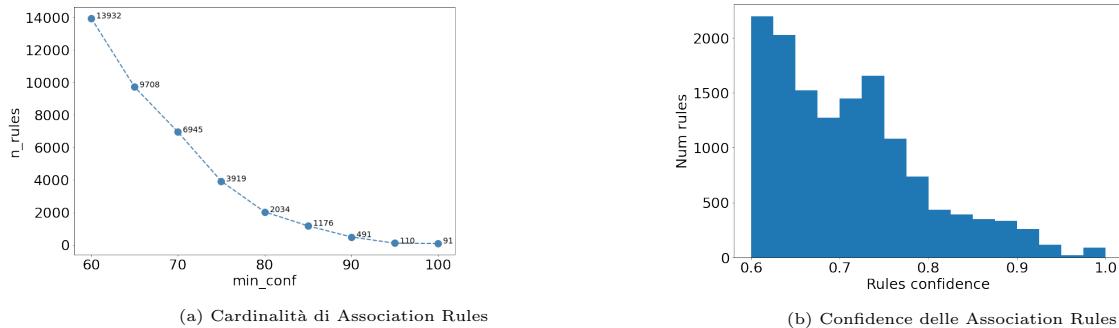


Figure 29

Per quanto riguarda il valore del lift si osserva in figura 30 che ci sono 2 picchi, di cui quello in corrispondenza al valore 1 dominante rispetto l'altro; ciò sta a significare che molte delle regole estratte saranno formate da valori pressoché statisticamente indipendenti tra di loro. Questo risultato è attribuibile allo sbilanciamento dei valori di molti attributi del dataset, che saranno quindi ampiamente diffusi indipendentemente dalle regole di associazione.

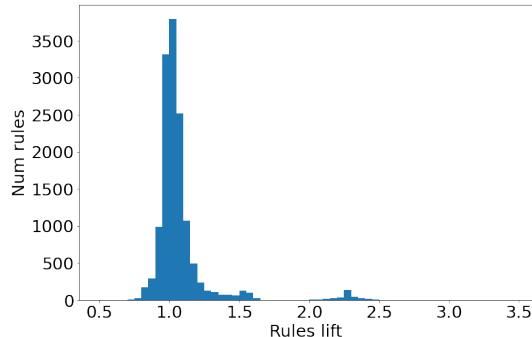


Figure 30: Distribuzione lift

Utilizzando l'algoritmo Apriori con parametri supp=10, zmin=3, conf=80 si ottengono 2034 regole di cui ne sono state selezionate alcune più interessanti, prendendo in cosiderazione i diversi conseguenti presenti nelle implicazioni.

- { TotalWorkingYears: [0; 10], StockOptionLevel: 0, NumCompaniesWorked: [0; 2], OverTime: No } \Rightarrow MaritalStatus: Single
Absolute Support: 134, Relative Support: 0.114, Confidence: 0.80, Lift: 2.478
- { MaritalStatus: Single, Attrition: No } \Rightarrow StockOptionLevel: 0
Absolute Support: 284, Relative Support: 0.241, Confidence: 1.0, Lift: 2.288
- { YearsInCurrentRole:[7; 10], TotalWorkingYears:[0; 10] } \Rightarrow NumCompaniesWorked:[0; 2]
Absolute Support: 122, Relative Support: 0.104, Confidence: 0.853, Lift: 1.423

- { YearsInCurrentRole: [0; 2], Satisfaction: [12; 14], NumCompaniesWorked: [0; 2] } \Rightarrow TotalWorkingYears[0; 10]
 Absolute Support: 134, Relative Support: 0.114, Confidence: 0.905, Lift: 1.423
- { YearsInCurrentRole: [0; 2], Satisfaction: [12; 14], TotalWorkingYears: [0; 10], Attrition: No } \Rightarrow OverTime: No
 Absolute Support: 138, Relative Support: 0.117, Confidence: 0.868, Lift: 1.21
- { Satisfaction: [15; 16], OverTime: No, Attrition: No } \Rightarrow PerformanceRating: 3
 Absolute Support: 164, Relative Support: 0.139, Confidence: 0.832, Lift: 1.112

Alcune di queste presentano lift che si aggira intorno ad 1 e sono relative a valori come PerformanceRating uguale a 3 e OverTime No, che occorrono oltre 800 volte su 1176 nel dataset perciò costituiscono regole non molto significative. Le altre association rules trovate invece, pur mantenendo una confidence maggiore all'80%, si vanno a posizionare nel secondo picco precedentemente citato.

5.3.1 Sostituzione Missing values

I missing values presenti nel dataset sono i seguenti: BusinessTravel, MonthlyIncome, Age, TrainingTimesLastYear, PerformanceRating, Gender. Al fine di trovare delle regole adatte che descrivessero i valori di questi attributi, è stato necessario ridurre la soglia minima di support in quanto alcuni valori occorrono con meno frequenza all'interno del dataset.

Per BusinessTravel sono state trovate regole relative al valore 'Travel_Rarely' (confidence 85%, lift 1.31, support 51), mentre non ne sono state trovate per 'Non_Travel' e 'Travel_Frequently' per valori di confidence maggiori di 60%, soglia oltre la quale non sarebbe utile andare a cercare. Tuttavia solo 2 dei 107 missing values sono sostituiti attraverso questa regola.

Per Gender sono state trovate sia le regole per il valore 'Male' che per 'Female', rispettivamente con confidence 80% (lift 1.4, support 50) e 61% (lift 1.59, support 38), grazie alle quali è stato possibile rimpiazzare 5 missing values su 59.

Per PerformanceRating, ad 11 su 138 missing values è stato possibile assegnare il valore 3 (confidence 90%, lift 1.2, support 58).

Per MonthlyIncome sono stati sostituiti 6 valori su 213 assegnandogli la classe MonthlyIncome[1009.0; 6244.0] (confidence 81%, lift 1.55, support 48).

Non sono state trovate invece association rules per Age e TrainingTimesLastYear.

5.3.2 Predizione dell'Attrition

Per quanto riguarda la predizione dell'Attrition, per il valore 'No' le association rules relative presentano una confidence alta ma un basso valore di lift: ciò fa concludere che non sono adatte per predire valori in quanto influenzate dallo sbilanciamento del dataset (984 No, 192 Si). Mentre per il valore 'Si' la situazione è opposta, con le regole trovate che presentano lift alta e confidence bassa. In particolare è riportata qui sotto la regola con lift e confidence maggiori (59% confidence, 3.6 lift, 36 support):

- { OverTime: Si, StockOptionLevel: 0, YearsInCurrentRole: [0; 2], TotalWorkingYears: [0; 10] } \Rightarrow Attrition: Si

Benchè il valore di lift sia alto, la confidence sotto il 60% non ci permette di affermare che sia una regola particolarmente significativa.