Università di Pisa

Data Mining project

Perché i lavoratori lasciano un'azienda?

Il caso IBM

A cura di

Erica Cau, Alfonso Ferraro, Simona Mazzarino, Federico Mazzoni

1 Introduzione

Il *dataset IBM HR* è una raccolta di dati creata da IBM (*International Business Machine Corporation*), azienda americana leader nel settore informatico, per studiare le ragioni che portano i propri dipendenti all'autolicenziamento. L'obiettivo della seguente indagine è dunque quello di analizzare il dataset per comprendere quali siano le variabili che più frequentemente influenzano la suddetta scelta.

Il progetto è suddiviso in quattro sezioni: *Data Understanding, Clustering*, *Classification* e *Association Rules Mining*. Nel paragrafo dedicato al *Data Understanding* prepareremo il dataset attraverso la *Data Semantics*, la distribuzione statistica dei dati, la *Data Quality* e infine la pulizia dei dati stessi.

Le parti successive saranno destinate all'esplorazione del dataset e alla comprensione del fenomeno dell'abbandono del posto di lavoro per mezzo di algoritmi di *clustering, association rules mining* e *classification*. L'ultima sezione sarà invece riservata alle considerazioni finali.

2 Data Understanding

Categorici	
Ordinali	**Nominali**
Education	Attrition
EnvironmentSatisfation	BusinessTravel
JobInvolvement	Department
JobSatisfaction	EducationField
PerformanceRating	Gender
RelationshipSatisfaction	JobRole
WorkLifeBalance	MaritalStatus
StockOptionLevel	Over18
JobLevel	OverTime

Numerici	
Continui	**Discreti**
Age	DailyRate
TotalWorkingYears	MonthlyRate
Hourly Rate	MonthlyIncome
YearsAtCompany	
DistanceFromHome	
NumCompaniesWorked	
PercentSalaryHike	
StandardHours	
TrainingTimeLastYear	
YearsInCurrentRole	
YearsSinceLastPromotion	
YearsWithCurrentManager	

Tabella 1: *Attributi categorici e numerici nel dataset*

Il dataset IBM HR è formato da *1176 record* (o oggetti) e *33 feature* (o attributi) di cui 18 categoriche e 15 numeriche. All'interno della colonna degli attributi nominali, nella tabella 1, alcune variabili, nello specifico *Attrition, Gender, Over18, OverTime*, sono classificate come binarie. L'attributo *Over18* presenta, inoltre, la caratteristica di possedere i seguenti valori: "Y", interpretato come *Yes* e valore "NaN".

2.1 Data semantic

Tra le *feature* più interessanti del dataset, figura sicuramente l'attributo binario *Attrition*, ovvero il tasso di abbandono della posizione lavorativa dalla IBM, a cui sono associati due valori: *Yes* e *No*.

Ad essa si ricollegano altri attributi volti a indagare il livello d'istruzione dei dipendenti, il loro campo di studi (*Education* e *EducationField*), il ruolo all'interno dell'azienda (*JobRole*) e il livello delle performance lavorative (*PerformanceRating*). Viene anche posta attenzione al tempo destinato alla formazione aziendale, agli anni di impiego sotto lo stesso manager (*YearsWithCurrentManager*) e agli anni nel ruolo attuale (*YearsInCurrentRole*).

Per ogni impiegato sono inoltre valutate diverse sfumature di soddisfazione, legate sia al rapporto con l'azienda che con il lavoro in sé (*JobSatisfaction* e *RelationshipSatisfaction*), sia rispetto all'ambiente lavorativo (*EnvironmentSatisfaction*).

feature	**Descrizione**
Age, Over18	Età dei dipendenti e maggiore età
Attrition	Abbandono della posizione lavorativa
BusinessTravel	Viaggi di lavoro
HourlyRate, DailyRate, MonthlyRate	Tariffa oraria, giornaliera e mensile
Department	Reparto aziendale di lavoro
DistanceFromHome	Distanza in KM dal domicilio
Education, EducationField	Livello d'istruzione e ambito di studi
EnvironmentSatisfaction	Gradimento dell'ambiente lavorativo misurato in scala numerica
Gender	Sesso del dipendente
JobInvolvement	Coinvolgimento nel lavoro misurato in scala numerica
JobLevel	Livello della posizione lavorativa misurato in scala numerica
JobSatisfaction	Gradimento del lavoro misurato in scala numerica
JobRole	Posizione lavorativa
MaritalStatus	Stato civile
MonthlyIncome, PercentSalaryHike	Stipendio mensile e Aumento salariale in percentuale
NumCompaniesWorked	Numero di aziende in cui il dipendente ha lavorato precedentemente
PerformanceRating	Valutazione delle prestazioni misurata in scala numerica
RelationshipSatisfaction	Gradimento del rapporto tra dipendente e azienda
StockOptionLevel	Piani di *Stock Option* offerti dall'azienda ai dipendenti come incentivo
TotalWorkingYears	Totale degli anni in cui il dipendente ha lavorato nel corso della sua vita
StandardHours, OverTime	Totale delle ore lavorative contrattuali e straordinari
TrainingTimeLastYear	Periodo di formazione nell'ultimo anno
WorklifeBilance	Equilibrio tra lavoro e vita privata misurato in scala numerica
YearsAtCompany	Anni di impiego alla IBM
YearsInCurrentRole	Totale degli anni in cui il dipendente ricopre la stessa posizione lavorativa
YearsSinceLastPromotion	Totale degli anni trascorsi dall'ultima promozione
YearsWithCurrentManager	Totale degli anni trascorsi sotto lo stesso dirigente

Tabella 2: *Descrizione degli attributi del dataset*

2.2 Distribuzione statistica delle *feature*

In questa sezione ci dedicheremo a un'analisi più approfondita delle distribuzioni statistiche delle *feature* più rilevanti, studiandole sia individualmente sia in correlazione tra loro, mentre in seguito discuteremo brevemente i risultati delle operazioni statistiche applicate alle *feature* categoriche e numeriche.

Attrition, la *feature* binaria da cui tutta la nostra analisi è partita, può assumere i valori *Yes* e *No*. Come si può osservare nella figura 1 la distribuzione è fortemente sbilanciata: è molto più elevato il numero di dipendenti che rimangono in azienda (83,67% sul totale) rispetto a quelli che decidono di lasciarla (16,33%). Una prima intuizione ci ha portato a correlarla con lo stipendio mensile (*MonthlyIncome*) e con il numero totale degli anni trascorsi dall'ultima promozione *YearsSinceLastPromotion*.

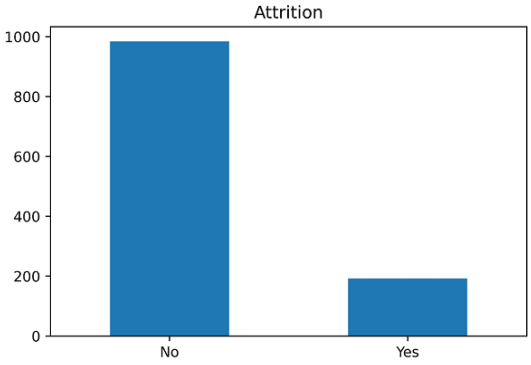
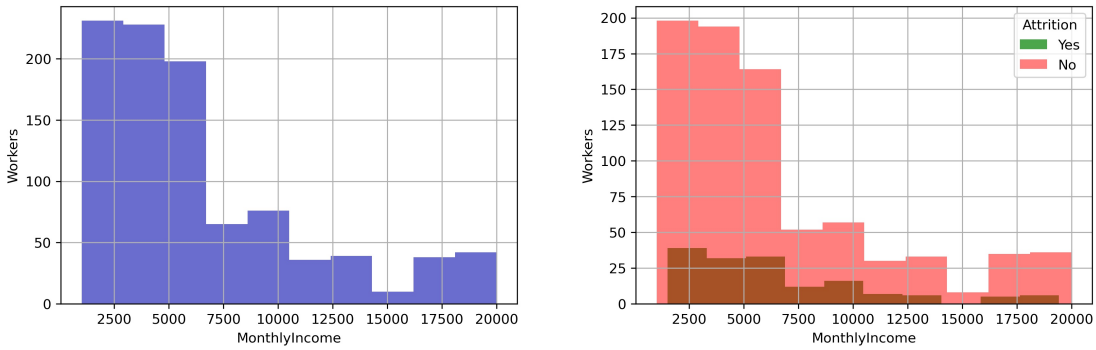


Figura 1: Distribuzione della *feature Attrition*

MonthlyIncome, lo stipendio mensile di ogni lavoratore, come possiamo vedere in figura 2a, ha una distruzione sbilanciata verso i valori bassi. Risulta quindi un numero molto elevato di lavoratori che percepisce uno stipendio in un range che va dai 2500$ ai 7500$ e un numero considerevolmente più basso di lavoratori con stipendi medio-alti rispetto ai valori presenti nel dataset. È tuttavia presente un leggero aumento nella fascia di dipendenti che percepiscono una paga mensile compresa tra 17000$ e 20000$.

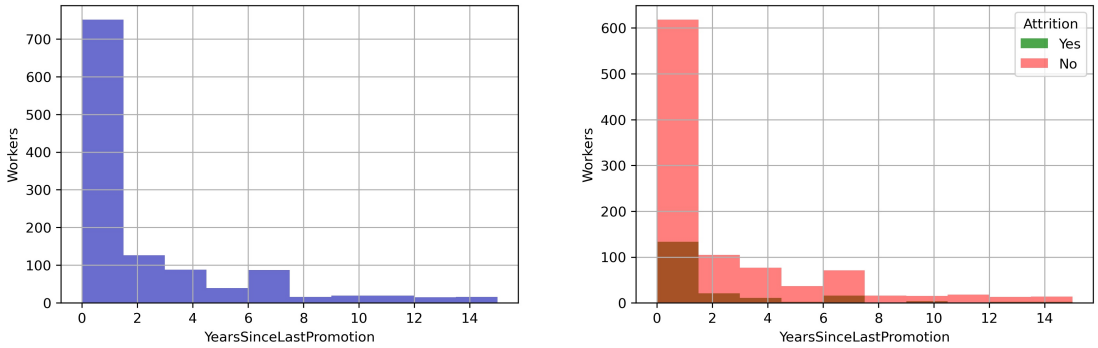


(a) **(b)**

Figura 2: Distribuzione statistica dello stipendio mensile e del tasso di *Attrition* in base allo stipendio

Come rappresentato in figura 2b, la maggior parte delle persone che hanno deciso di lasciare il proprio posto di lavoro otteneva a fine mese uno stipendio compreso in un range basso. Tuttavia, anche in questo grafico, come in quello precedente, si nota un lieve incremento di *Attrition* in corrispondenza di stipendi elevati: in questo contesto, si può ipotizzare che la crescita dei valori positivi di *Attrition* sia dovuta a possibili pensionamenti.

YearsSinceLastPromotion è la *feature* che descrive gli anni trascorsi dall'ultima promozione. Anche la sua distribuzione, come quella della *feature* precedente, è fortemente sbilanciata verso lo zero (cfr. Figura 3a). È interessante notare come anche i valori dell'*Attrition* sono più elevati in corrispondenza dei valori 0 e 1; ciò potrebbe significare che molti lavoratori decidono di cambiare azienda dopo il primo periodo di formazione, oppure, dopo pochi anni dall'assunzione.



(a) **(b)**

Figura 3: Distribuzione statistica degli anni dall'ultima promozione e del tasso di *Attrition* in base allo stipendio

JobRole è la posizione lavorativa di ciascun dipendente. Come viene evidenziato nella Figura 4, i lavoratori che tendono maggiormente a lasciare l'azienda sono *Laboratory Technicians*, seguiti dai *Research Scientists* e *Sales Executives*. Vi sono, invece, pochissimi dipendenti che lasciano le posizioni di *Manager* e *Healthcare Representative* e addirittura nessuno che lascia la posizione di *Research Director*.

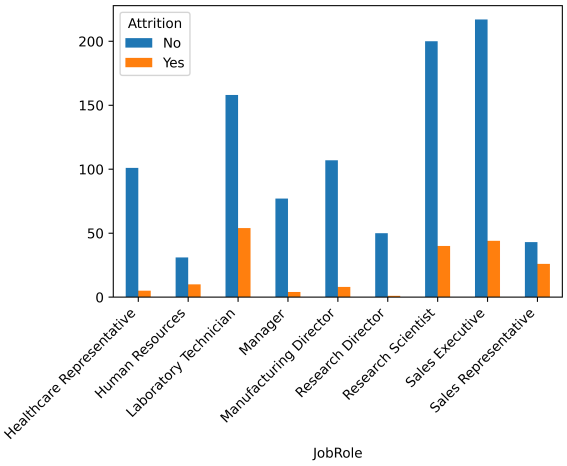
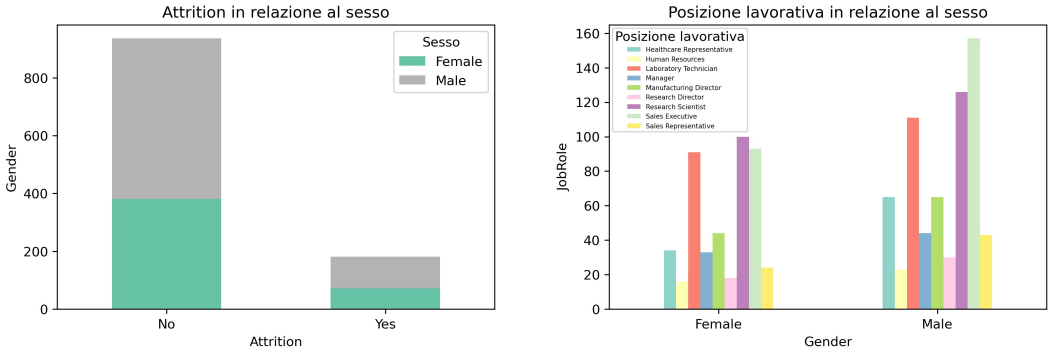


Figura 4: Distribuzione della *feature Attrition* correlata al *JobRole*

Gender è l'attributo riguardante il sesso dei dipendenti. Dal grafico in Figura 5a notiamo che la quantità di persone che decidono di lasciare l'azienda è più equilibrata rispetto a quella osservata in *JobRole*, sebbene in proporzione siano più gli uomini a prendere questa decisione rispetto alle donne. Abbiamo inoltre deciso di visualizzare la distribuzione del *Gender* in base alla posizione lavorativa: anche in questo ambito le professioni appaiono più equamente distribuite tra i sessi (cfr. Figura 5b). Per le donne spiccano i valori associati a *Laboratory Technician* e a *Research Scientists*: in proporzione, infatti, risulta una maggior presenza di donne in questi settori rispetto agli uomini. Per i dipendenti di sesso maschile, invece, saltano all'occhio le professioni di *Sales Executive* e *Healthcare Representative*.



(a) **(b)**

Figura 5: Distribuzione dell'*Attrition* in relazione al sesso dei lavoratori e distribuzione dei diversi ruoli svolti in base all'attributo *Gender*

Nella seguente tabella (Tabella 3) abbiamo riportato le analisi statistiche applicate ai vari attributi numerici del dataset. Da questa si possono evincere alcuni valori chiave per lo studio dei dati, come ad esempio, l'età media dei lavoratori (37 anni circa) e lo stipendio mensile medio, minimo e massimo (in ordine 6566$, 1009$ e 19999$).

Inoltre, tenendo in considerazione le medie delle *feature* relative a *Environment Satisfaction*, *Job Satisfaction* e *Relationship Satisfaction*, si nota che i dipendenti IBM sono soddisfatti sia del proprio lavoro che dell'ambiente lavorativo aziendale. Infine, il livello medio di *Education* tende a 3: tendenzialmente i lavoratori posseggono una laurea triennale (*Bachelor's degree*).

features	**count**	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**
Age	1000.0	37.199000	9.015802	18.0	30.00	36.0	43.00	60.0
DailyRate	1176.0	803.650510	406.683045	102.0	460.50	804.0	1169.00	1499.0
DistanceFromHome	1176.0	9.210034	8.097024	1.0	2.00	7.0	14.00	29.0
Education	1176.0	2.884354	1.016574	1.0	2.00	3.0	4.00	5.0
EnvironmentSatisfaction	1176.0	2.715986	1.088876	1.0	2.00	3.0	4.00	4.0
HourlyRate	1176.0	66.299320	20.266116	30.0	49.00	66.0	84.00	100.0
JobInvolvement	1176.0	2.735544	0.716228	1.0	2.00	3.0	3.00	4.0
JobLevel	1176.0	2.021259	1.069686	1.0	1.00	2.0	3.00	5.0
JobSatisfaction	1176.0	2.702381	1.101578	1.0	2.00	3.0	4.00	4.0
MonthlyIncome	963.0	6565.946002	4710.625603	1009.0	2969.00	4969.0	8585.00	19999.0
MonthlyRate	1176.0	14395.836735	7111.845106	2097.0	8227.25	14434.0	20489.25	26999.0
NumCompaniesWorked	1176.0	2.663265	2.491287	0.0	1.00	2.0	4.00	9.0
PercentSalaryHike	1176.0	15.176871	3.623941	11.0	12.00	14.0	18.00	25.0
PerformanceRating	1038.0	3.152216	0.359403	3.0	3.00	3.0	3.00	4.0
RelationshipSatisfaction	1176.0	2.702381	1.092268	1.0	2.00	3.0	4.00	4.0
StandardHours	606.0	80.000000	0.000000	80.0	80.00	80.0	80.00	80.0
StockOptionLevel	1176.0	0.783163	0.851385	0.0	0.00	1.0	1.00	3.0
TotalWorkingYears	1176.0	11.019558	7.694848	0.0	6.00	10.0	15.00	40.0
TrainingTimesLastYear	943.0	2.827147	1.273120	0.0	2.00	3.0	3.00	6.0
WorkLifeBalance	1176.0	2.755952	0.707984	1.0	2.00	3.0	3.00	4.0
YearsAtCompany	1116.0	6.926523	6.063193	0.0	3.00	5.0	9.00	40.0
YearsInCurrentRole	1176.0	4.188776	3.637405	0.0	2.00	3.0	7.00	18.0
YearsSinceLastPromotion	1176.0	2.171769	3.189785	0.0	0.00	1.0	3.00	15.0
YearsWithCurrManager	1176.0	4.107993	3.601097	0.0	2.00	3.0	7.00	17.0

Tabella 3: *Statistiche degli attributi numerici con totale dei valori, media, deviazione standard, valore minimo e massimo, quartili (25%, 50%, 75%).*

Passando all'analisi dei risultati in tabella 4, è possibile definire una migliore profilazione dei dipendenti IBM: sono prevalentemente di sesso maschile (il 56,46% sul totale dei dipendenti), sposati, che hanno svolto studi legati alle scienze naturali o alle discipline sanitarie. All'interno dell'azienda, i dipendenti lavoravano prevalentemente come *Sales Executive*, ricercatori o tecnici di laboratorio. Si può infine evidenziare come l'81,29% non svolga viaggi di lavoro o li effettui raramente.

features	**Righe non vuote**	**Valori unici**	**1° valore più frequente (moda)**	**2° valore più frequente.**	**3° valore più frequente**
Attrition	1176	2	No (984)	Yes (192)	/
BusinessTravel	1069	3	Travel_Rarely (764)	Travel_Frequently (192)	Non-Travel (113)
Department	1176	3	Research & Development (769)	Sales (361)	Human Resources (46)
EducationField	1176	6	Life Sciences (489)	Medical (370)	Marketing (125)
Gender	1117	2	Male (664)	Female (453)	/
JobRole	1176	9	Sales Executive (261)	Research Scientist (240)	Laboratory Technician (212)
MaritalStatus	1176	3	Married (542)	Single (383)	Divorced (251)
Over18	804	1	Y (804)	/	/
OverTime	1176	2	No (838)	Yes (338)	/

Tabella 4: *Descrizione statistica degli attributi categorici nominali con frequenza dei primi tre valori più ricorrenti.*

3 Valutazione della qualità dei dati

3.1 Valori mancanti

Sono stati rilevati valori mancanti per gli attributi: *Age*, *BusinessTravel*, *Gender*, *Over18*, *MonthlyIncome*, *PerformanceRating*, *StandardHours*, *TrainingTimesLastYear* e *YearsAtCompany*.

Non si è proceduto alla correzione di *Over18*, *StandardHours* e *PerformanceRating*: i valori presenti non permettono di effettuare una stima soddisfacente delle controparti mancanti (cfr. *Attributi problematici* nella Tab. 5 e, più sotto, le considerazioni su *PerformanceRating*).

I valori mancati di *Age* e *TrainingTimesLastYear* sono stati sostituiti con la media arrotondata all'intero più vicino, mentre nel caso di *BusinessTravel* e *Gender* con la moda. In tutti e quattro i casi, i valori sono stati calcolati tenendo come riferimento il rispettivo *JobRole*, nel tentativo di offrire una stima quanto più accurata possibile. La media delle età dei lavoratori *Sales Executive*, ad esempio, è 38 anni, mentre dei *Research Director* 40.

Attributi corretti		Attributi problematici	
feature	**Valori mancanti** (%)	***feature***	**Valori**
Age	176 (15%)	**Over18**	*NaN* 372 (32%) *Y* 804 (68%)
Gender	59 (5%)	**StandardHours**	*NaN* 570 (48%) *80.0* 606 (52%)
BusinessTravel	107 (9%)	**PerformanceRating**	*NaN* 138 (18%) *3.0* 808 (69%) *4.0* 158 (13%)

Tabella 5: Valori mancanti: attributi corretti e problematici

3.2 Errori nel dataset

In *MonthlyIncome* e *YearsAtCompany* ai valori mancanti si accompagnano svariati errori di correttezza semantica.

Il confronto fra *YearsAtCompany* e gli analoghi attributi *TotalWorkingYears*, *YearsInCurrentRole*, *YearsSinceLastPromotion* e *YearsWithCurrManager*, o con *Age* e *NumCompaniesWorked*, mostra in molti casi valori contraddittori[1].

YearsInCurrentRole è l'unico attributo mantenuto in quanto più relazionato ad attrition, e coerente una volta osservato con YearsWithCurrManager. Pochi outliers.

Singolarmente, *YearsAtCompany* non mostra poi alcuna correlazione con *Attrition*. Si è notato che rimuovendo gli *outlier* si ha un lieve aumento della correlazione, che rimane tuttavia al di sotto di quella degli altri attributi. Si è quindi scelto di mantenere come solo parametro riferito agli anni *YearsInCurrentRole*. Per i fini del presente *paper* è infatti il più interessante, mostrando una discreta correlazione con *Attrition*, e la quasi perfetta corrispondenza fra suoi i valori e quelli presenti in *YearsWithCurrManager* rende plausibile supporne una buona correttezza.

MonthlyIncome unico valore di Rate scelto

Analogamente si è proceduto con *MonthlyIncome*, *HourlyRate*, *DailyRate* e *MonthlyRate*, tutti quanti contraddittori fra loro. Rimuovendo gli outlier e sostituendo i valori mancanti (con la metodologia di cui sopra), *MonthlyIncome* ha mostrato una correlazione con *Attrition* ed è stato l'unico attributo mantenuto.

3.3 Rimozione degli *outlier*

Per l'analisi di alcuni degli attributi del paragrafo precedente si è fatto ricorso alla rimozione degli *outlier*, individuati con l'uso di *boxplot*. Di ciascun parametro è stato poi calcolato l'*IQR* e si è proceduto alla rimozione dei valori non appartenenti al *range interquartile*.

[1] Ad esempio, *TotalWorkingYears* dovrebbe essere sempre superiore a *YearsAtCompany* e, trattandosi IBM di una compagnia statunitense, la differenza fra qualsiasi parametro riferito agli anni lavorativi e *Age* non dovrebbe mai essere superiore a 14.

La stessa operazione è stata effettuata per *YearsInCurrentRole*. I (pochi) *outlier* rilevati confermano ulteriormente la validità di questo attributo. Nella Fig. 6 gli *outlier* di *YearsInCurrentRole* a confronto con quelli di *YearsAtCompany*.

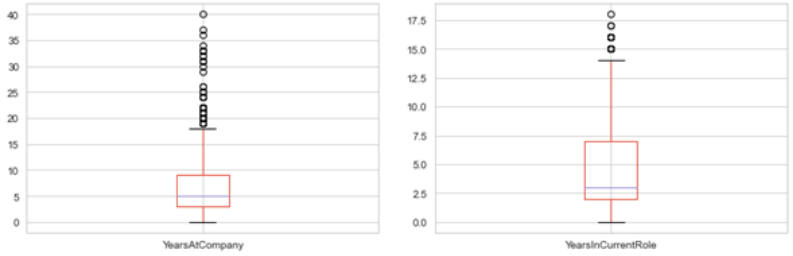


Figura 6: Boxplot per *YearsAtCompany* e *YearsInCurrentRole*

3.4 Creazione di attributi

Per ridurre dimensione dataset, si sono unite varie features creandone di nuove usando le medie. Nel caso di company involvement si sono riscalati i valori.

Per agevolare l'analisi del dataset, sono stati creati due nuovi attributi in luogo di altri già esistenti.

- ***GeneralEmployeeSatisfaction***: include la media approssimata all'intero più vicino dei valori riguardanti la soddisfazione del dipendente: *EnvironmentSatisfaction*, *JobSatisfaction*, *RelationshipSatisfaction*;
- ***CompanyInvolvement***: include la media approssimata all'intero più vicino dei valori degli attributi indici del coinvolgimento del dipendente con le dinamiche aziendali: *StockOptionLevel* e *JobInvolvement*. Si noti che il primo parametro era in scala 0-3, il secondo 1-4: i valori di *StockOptionLevel* sono stati dunque convertiti.

3.5 Trasformazione dei valori

I valori di alcuni attributi di tipo stringa sono stati convertiti in numerici[2]. In particolare:

- ***Attrition***: NO → 0, YES → 1;
- ***Business Travel***: Non_Travel → 0, Travel_Rarely → 1, Travel_Frequently → 2.
- ***Department***: Human Resources → 0, Research & Development → 1, Sales → 2;
- ***EducationField***: Human Resources → 0, Life Sciences → 1, Marketing → 2, Medical → 3, Other → 4, Technical Degree → 5;
- ***Gender***: Female → 0, Male → 1;
- ***JobRole***: Healtcare Rappresentative → 0, Human Resources → 1, Laboratory Technician → 2, Manager → 3, Manufacturing Director → 4, Research Director → 5, Research Scientist → 6, Sales Executive → 7, Sales Rappresentative → 8;
- ***MaritalStatus***: Divorced → 0, Married → 1, Single → 2;
- ***OverTime***: No → 0, Yes → 1.

3.6 Rimozione di *feature*

La correzione degli errori e la conversione dei sopracitati valori (in particolare dell'attributo-chiave *Attrition*) permettono di computare una prima *matrice di correlazione*. Da essa, possiamo notare quali siano gli attributi più importanti e quali invece non abbiano alcuna particolare relazione con *Attrition*.

Alla luce dei rilevamenti finora segnalati, dal dataset sono stati eliminati i seguiti attributi, oltre a quelli già citati:

- ***StandardHours*** e ***Over18***: i valori non mancanti sono identici l'un l'altro, non offrendo alcuna informazione utile. *Over18* risulta inoltre ridondante con *Age*;

[2]Per praticità, nel dataframe tali attributi sono stati siglati con * (es. *Attrition**).

- ***PerformanceRating***: rispetto ai due casi precedenti, i valori mancanti sono relativamente più contenuti e a livello intuitivo il parametro sarebbe stato potenzialmente molto utile per l'analisi di *Attrition*. Tuttavia, l'ampia presenta di "3" a scapito di "1" e "2" (che non hanno alcuna occorrenza) inficia gravemente i possibili utilizzi di questo parametro e, a nostro avviso, la sua generale validità. È stato pertanto scartato;
- ***Education***, ***BusinessTravel***, ***TrainingTimesLastYear***, ***NumCompaniesWorked***, ***PercentSalaryHike***: scarsa o nulla correlazione con *Attrition*;
- ***JobRole***: il parametro è stato utile per correggere i dati mancanti, ma per le successive analisi risulta ridondante, con la presenza del più compatto *Department*;
- ***EducationField***: nessuna correlazione con *Attrition*. Per questo parametro abbiamo verificato se l'eventuale discrasia fra *EducationField* e *Department* possa essere una possibile causa di *Attrition*, ottenendo risposta negativa.

A seguito della nostra operazione di *data cleaning* abbiamo ottenuto un dataset di 1029 oggetti descritti da 13 attributi. Di seguito la relativa matrice di correlazione (sono esclusi *Department* e *MaritalStatus*, parametri non binari originariamente categorici).

Calcolata con coefficiente di Pearson

SOMMA(x-X)(y-Y)/(n-1)sxsy

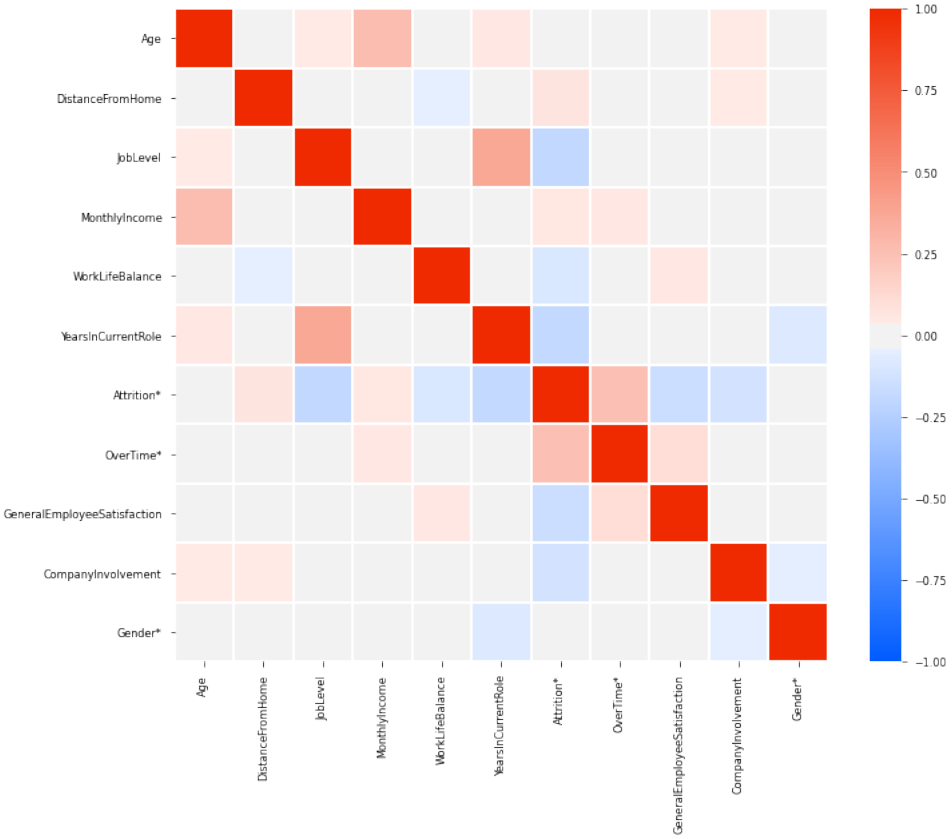


Figura 7: Matrice di correlazione dopo l'operazione di *data cleaning*

4 Clustering

A seguito del *data cleaning* si è proceduto al clustering del dataset, utilizzando come attributi di riferimento quelli che maggiormente descrivono la situazione del lavoratore all'interno un'azienda e possono determinarne l'*Attrition*: *MonthlyIncome*, *DistanceFromHome*, *JobLevel* e *OverTime*. Tutti e quattro i parametri hanno inoltre mostrato in una qualche misura correlazione con *Attrition*, come visto nella matrice. Lo scopo dell'operazione è stato suddividere i dipendenti in cluster con attributi correlati l'un l'altro. Sono stati utilizzati tre tipi di algoritmi di clustering: il *K-Means*, il *DBSCAN* e il *Gerarchico* (nelle sue varie declinazioni). Prima dell'applicazione di ciascun algoritmo è stato necessario effettuare lo *scaling* dei dati (attraverso il *MinMaxScaler*). Successivamente, sono stati valutati i parametri ottimali e, infine, la bontà dei risultati ottenuti. v' = v-minA/maxA-minA

4.1 *K-Means*

La corretta applicazione dell'algoritmo *K-Means* è legata alla scelta del *parametro K*, che rappresenta il numero di cluster in cui il dataset verrà suddiviso. Per rilevare il corretto valore di K abbiamo utilizzato il *Knee Method*, dopo aver rappresentato graficamente il possibile *SSE (Sum of Squared Errors)* per differenti quantità di cluster comprese fra 2 e 50. Dall'analisi della figura 8, si è notato che l'ideale valore di K è 4. Si è quindi calcolato il relativo *coefficiente di Silhouette*, pari a 0.397. Con valori di K più grandi si è notata una diminuzione del valore della Silhouette. Con K pari a 3 l'aumenta del valore è minimo (0.407), mentre avere soltanto due cluster (*Silhouette* = 0.518) sarebbe poco utile per i nostri scopi.

∑∑x in Ci (M-x)2

b-a/max(a,b)

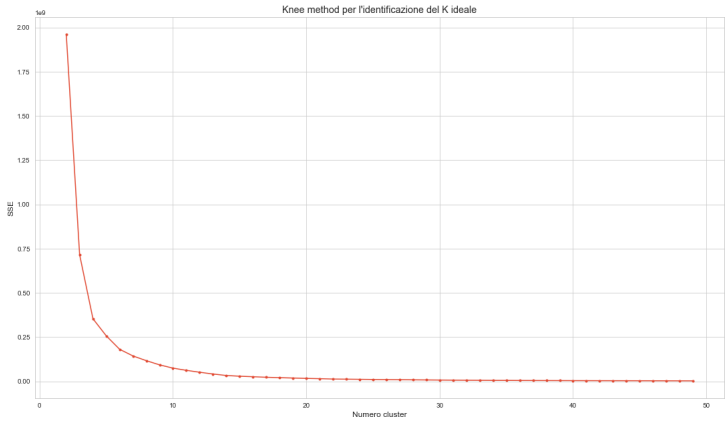


Figura 8: SSE con K da 2 a 50

Di seguito il grafico delle *Parallel Coordinates* dei quattro cluster individuati e la tabella che quantifica la popolazione di ciascun cluster. Tutti i cluster sono accomunati da un simile valore medio per *MonthlyIncome* e – a eccezione del Cluster 2 – assenza di *OverTime*. Sempre il Cluster 2 spicca per il *JobLevel* più alto della media. Valori più alti della media si rilevano anche nel Cluster 4 per la *DistanceFromHome*, invece molto bassa per i lavoratori del Cluster 1.

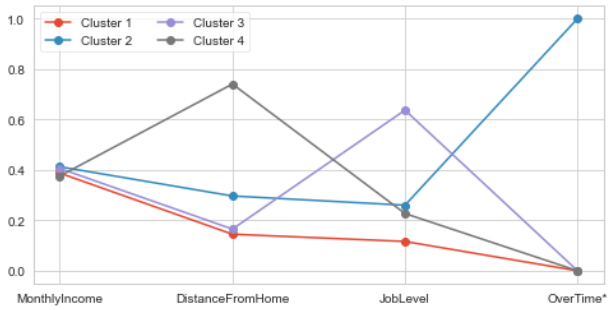


Figura 9: *Parallel Coordinates* dei cluster

Cluster	Numero dipendenti nel cluster
0	417
1	301
2	137
3	174

Tabella 6: Distribuzione dei dipendenti nei cluster

Si è poi passati a osservare la distribuzione dei vari parametri, a cominciare da *Attrition*. Essa appare pressoché identica fra i Cluster 1 e 4, con una lieve diminuzione nel Cluster 3 e un significativo aumento nel Cluster 2 dovuto, probabilmente, alla presenza di *OverTime* che abbiamo rilevato. I Cluster 2 e 4 contengono inoltre una relativa alta concentrazione di dipendenti con bassa *WorkLifeBalance*.

Una possibile spiegazione della bassa concentrazione di *Attrition* nel Cluster 3 – già notato per l'alto *JobLevel* - è data dalla distribuzione di *YearsInCurrentRole*: oltre tre quinti dei dipendenti sono infatti in azienda da più di sei anni. Di contro, i dipendenti del Cluster 1, oltre ad avere un *JobLevel* più basso, sono anche in azienda da meno tempo.

I valori di *WorkLifeBalance* e *DistanceFromHome* del Cluster 4 non sembrano essere motivo di *Attrition*, a discapito di quanto si potrebbe pensare intuitivamente (e quanto potrebbe emergere dalla matrice di correlazione). Essi sono forse mitigati dalla concentrazione sopra la media di dipendenti con alto livello di *CompanyInvolvement*.

Attributi non direttamente legati alla vita aziendale dell'impiegato, come *MaritalStatus* o *Gender*, risultano invece equamente distribuiti e non degni di nota.

Complessivamente, il metodo *K-Means* è riuscito ad analizzare il dataset in modo significativo, mettendo in evidenza alcuni attributi che possono influenzare (in positivo o in negativo) l'*Attrition* dei vari dipendenti.

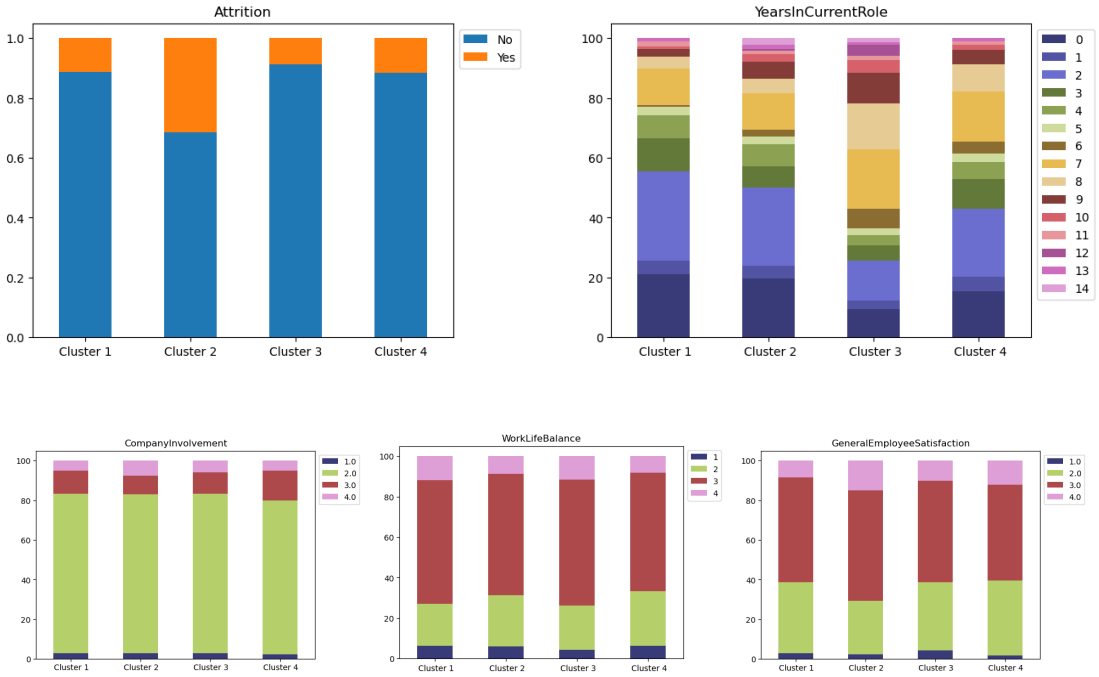


Figura 10: Distribuzione degli attributi *Attrition*, *YearsInCurrentRole*, *CompanyInvolvement*, *WorkLifeBalance*.*GeneralEmployeeSatisfaction* all'interno dei quattro cluster ottenuti

4.2 *DBSCAN*

Il metodo di clustering *DBSCAN* si è dimostrato inefficace per il nostro dataset, fortemente sbilanciato con un'alta concentrazione di valori con *Attrition* negativa. L'assunto iniziale – che l'*Attrition* sia determinata da un insieme di più attributi correlati l'un l'altro – sottintende poi un insieme di dati a più dimensioni. Le varie relazioni che intercorrono fra i vari parametri, di cui si è accennato sopra analizzando i risultati del *K-Means*, non possono essere correttamente visualizzate dal *DBSCAN*, che opera in due dimensioni. Dopo aver determinato 0.25 come valore ottimale di *EPS* (cioè il raggio dell'intorno di ciascun punto) ricorrendo al *Knee Method* (Fig.11), si è verificato che l'algoritmo genera

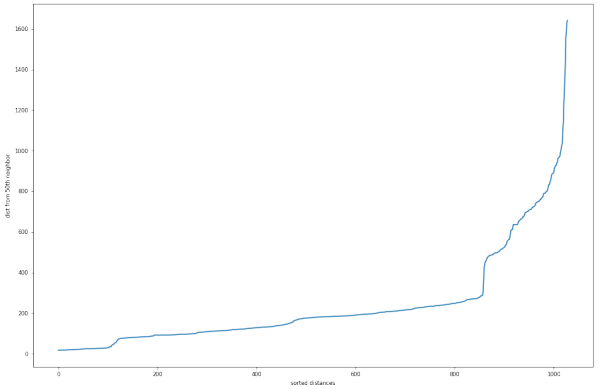


Figura 11: Grafico *Knee-method*

il modello con il migliore valore di Silhouette (0.26) fissando *minPts* (ovvero il numero minimo di punti che un punto deve avere nel suo intorno per essere definito *core*) pari a 4.

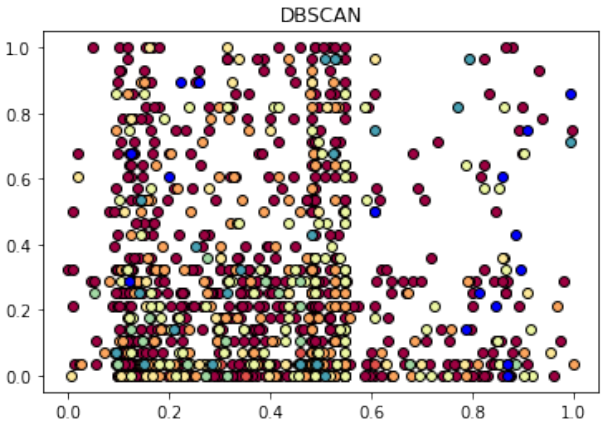


Figura 12: *DBSCAN*

Cluster	Numero dipendenti
Cluster 1	663
Cluster 2	7
Cluster 3	143
Cluster 4	44
Cluster 5	116
Cluster 6	19
Cluster 7	21
Rumore	16

Tabella 7: Distribuzione dei dipendenti nei cluster

Nello spazio sono tuttavia visualizzati dei punti senza alcuna apparente relazione fra loro (Fig. 12), divisi in otto diversi cluster, di cui uno riservato al rumore. Tali cluster sono molto più sbilanciati di quelli analizzati con il *K-Means*: abbiamo infatti un cluster eccessivamente grande, due medi e gli altri molto piccoli. Analoghi risultati sono stati ottenuti con altri valori di *minPts*.

Il basso valore della *Silhouette*, i limiti intrinseci al metodo e lo sbilanciamento dei risultati forniti dal modello fanno propendere contro l'utilizzo del *DBSCAN*.

4.3 Clustering gerarchico

Il terzo criterio di clustering sperimentato è quello gerarchico, che a sua volta ammette più possibili applicazioni: sono stati testati alberi gerarchici generati con l'algoritmo *Single Link* (o *MIN*), *Complete* (o *MAX*) e *Average*. In tutti i casi è stato necessario impostare un diverso *threshold* per ottenere un certo numero di cluster. Nel nostro caso, dovendo confrontare i risultati del metodo gerarchico con quelli ottenuti mediante *K-Means*, abbiamo cercato di ottenere sempre quattro cluster.

Metodo	Grandezza cluster	Threshold	Silhouette
Complete	69, 232, 180, 548	1.4	0.338
Average	726, 293, 8, 2	0.85	0.378
Single	728, 298, 2, 1	0.28	0.440

Tabella 8: Differenti parametri impostati per il clustering gerarchico

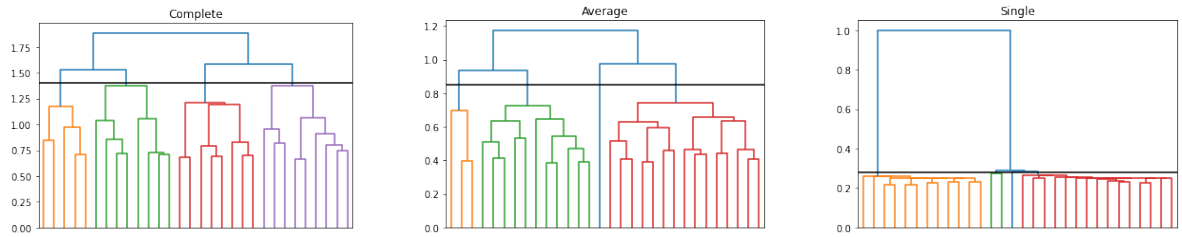


Figura 13: Cluster gerarchici ottenuti con i metodi *Complete*, *Average* e *Single link*

I risultati sono stati soddisfacenti solo con il metodo *Complete*. Negli altri casi i cluster risultano sbilanciati, con due cluster contenenti un numero molto basso di oggetti, un cluster eccessivamente grande e un cluster medio. A questo fenomeno vi sono due possibili spiegazioni.

Da una parte, poiché simili risultati sono stati ottenuti anche cercando differenti quantità di cluster, il problema potrebbe essere legato allo stesso dataset: l'alta quantità di rumore avrebbe portato gli algoritmi a costruire delle enormi "catene" incapaci di rilevare validi raggruppamenti per i diversi oggetti. Un'altra possibile spiegazione è che effettivamente sia preferibile avere due grandi cluster, situazione a cui *de facto* hanno portato i due metodi gerarchici. Il clustering generato dal metodo *Single Link* mostra poi un alto valore di *Silhouette*, coerentemente con quanto notato testando il *K-Means* impostando K pari a 2. Operare con solo due cluster, però, impedirebbe analisi più dettagliate dei diversi fattori che influenzano *Attrition*.

I risultati ottenuti dal metodo *Complete*, comunque, non confermano quelli del *K-Means*. La distribuzione degli oggetti è, nel complesso, diversa: nel *K-Means*, ad esempio, avevamo un solo cluster con ampia concentrazione di impiegati con *OverTime*.

4.4 Conclusioni sul clustering

Dalla nostra analisi l'algoritmo di clustering più efficace per l'analisi del dataset si è dimostrato il *K-Means*. Lo sbilanciamento del dataset e la forte presenza di rumore hanno reso impossibile applicare con risultati significativi i metodi *DBSCAN*, *Gerachico Single-Link* e *Gerarchico Average*. Risultati migliori sono stati ottenuti dal *Gerarchico Complete*, la cui Silhouette è tuttavia più bassa di quella del *K-Means*.

Algoritmo	**Silhouette**
K-Means	0.397
DBSCAN	0.26
Gerarchico (*Complete*)	0.338

Tabella 9: Tabella riassuntiva sugli algoritmi di clustering applicati e migliori valori della *Silhouette*

5 Classificazione

In questa parte si è classificato il dataset, al fine di predire in quali casi un lavoratore possa plausibilmente lasciare l'azienda o no (in termini più specifici, quale possa essere il valore della sua *Attrition*). Il metodo di classificazione *Decision Tree* è quello che più di tutti fornisce risultati facilmente leggibili ed è stato quindi scelto per primo. Il carattere fortemente sbilanciato del dataset, in particolare in riferimento all'*Attrition*, ha tuttavia, portato a risultati poco significativi.

KNN non crea modello, ma compara un record da etichettare con i suoi k records più vicini e assegna la classe di maggioranza o quella pesata. Algoritmo molto costoso, ma utile per una classificazione semplice.

Per ottenere un dataset più bilanciato, abbiamo allora applicato due tecniche di *sampling*: il *Random Undersampling*, che rimuove oggetti della classe maggioritaria (nel nostro caso *Attrition* pari a *No*) e la *Synthetic Minority Oversampling Technique* (*SMOTE*), che genera nuovi oggetti della classe in minoranza (con *Attrition* pari a *Yes*) osservando i preesistenti oggetti più simili.

Algoritmo che prevede la creazione di N decision tree che studiano parti random del dataset. La classe assegnata è la moda di tutti i DT. Utile per evitare l'overfitting del modello.

Successivamente, abbiamo applicato ulteriori algoritmi di classificazione al dataset, per osservare un possibile aumento dell'accuratezza. Nello specifico, sono stati applicati l'algoritmo *Random Forest* e *K-Nearest Neighbors* (KNN).

5.1 Preparazione del dataset

Il dataset in nostro possesso era già diviso in due parti: il *training set* e il *test set*. Dopo aver proceduto all'eliminazione dei valori mancanti in entrambi (con le procedure di cui abbiamo parlato sopra), abbiamo suddiviso il dataset di *training*: il 70% è stato destinato effettivamente al *training*, mentre il restante 30% ha formato il *validation set*, necessario per confrontare fra loro tutti i modelli generati dall'addestramento sul *training* al fine di trovare i parametri ottimali per i classificatori.

Validation non reinserito nel Training

Le tecniche di *Oversampling* e *Undersampling* sopra menzionate sono state applicate solo al *training set* in senso stretto.

5.2 Ottimizzazione dei parametri

Per quanto riguarda i metodi di *Decision Tree*, l'ottimizzazione dei parametri è stata eseguita applicando l'algoritmo di *random search* al *training set* originale, sia prima che dopo l'applicazione dello *SMOTE* o del *Random Undersampling*.

L'algoritmo ha addestrato diversi classificatori, testando di volta in volta diverse combinazioni dei parametri *Max Depth* (profondità massima dell'albero), *Min Sample Split* (numero minimo di oggetti per procedere a una ramificazione) e *Min Sample Leaf* (numero minimo di oggetti per foglia), oltre che possibili misure di impurità (indice di Gini o entropia).

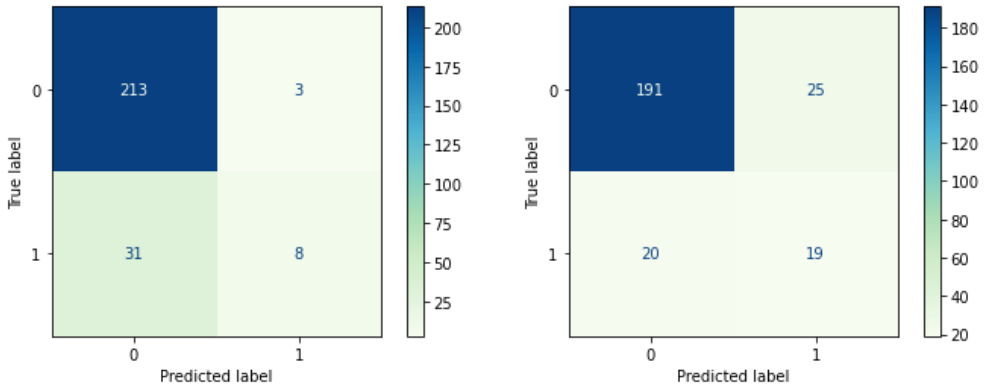
Nella tabella 10 riportiamo ciascun parametro dei vari modelli e i relativi risultati: l'accuratezza, il punteggio F1 medio (ovvero, la media della media armonica di *precision* e *recall* per i due valori di *Attrition*) e la curva di ROC (rapporto fra FP e TP). Tali risultati sono riportati sia per il *training set* che per il *validation set*.

	Modello 1	**Modello 2**	**Modello 3**	**Modello 4**	**Modello 5**	**Modello 6**
Sampling	Nessuno	Nessuno	*SMOTE*	*SMOTE*	*Random Undersampling*	*Random Undersampling*
Criterion	Gini	Gini	Entropia	Gini	Gini	Gini
Min Sample Split	29	15	31	4	6	23
Min Sample Leaf	34	25	12	5	9	5
Max Depth	4	12	17	13	17	4
Training Accuracy	0.851	0.854	0.895	0.931	0.848	0.791
Training F1-Score	1.24	1.35	1.79	1.86	1.70	1.58
Training AUC-ROC	0.594	0.645	0.895	0.931	0.848	0.791
Validation Accuracy	0.845	0.851	0.816	0.764	0.693	0.725
Validation F1-Score	1.11	1.24	1.17	1.17	1.19	1.26
Validation AUC-ROC	0.554	0.596	0.575	0.590	0.654	0.697

Tabella 10: Parametri, accuratezza, *F1-score* e *AUC-ROC* risultanti dalla *RandomGridSearch* con diverse tecniche di *sampling*

Nei modelli generati dall'algoritmo *Decision Tree*, in media i risultati ottenuti con il *training set* originale hanno valori di accuratezza di 85% (Modelli 1 e 2). Si è notato un sensibile miglioramento applicando al *training set* lo *SMOTE* (Modelli 3 e 4). Con il *Random Undersampling* non si sono ottenuti invece miglioramenti (Modello 5) e, in taluni casi (Modello 6), si è anzi constatata una diminuzione dell'accuratezza.

In tutti i modelli addestrati con *SMOTE* o *Random Undersampling* si nota una diminuzione dell'accuratezza sul *validation set*. In fase di test i risultati dell'accuratezza (misurata sia tramite *cross validation* impostata a 10, sia tramite *accuracy score*) hanno confermato questo fenomeno: i modelli ottenuti con tecniche di *sampling* soffrono così di *overfitting*.



(a) **(b)**

Figura 14: Matrici di confusione ottenute applicando sul *test set* il (a) *Modello 1* e (b) *Modello 4*. Si può notare la difficoltà dei modelli nel riconoscimento corretto dei dipendenti con *Attrition = 1*, ovvero i lavoratori che abbandonano la IBM.

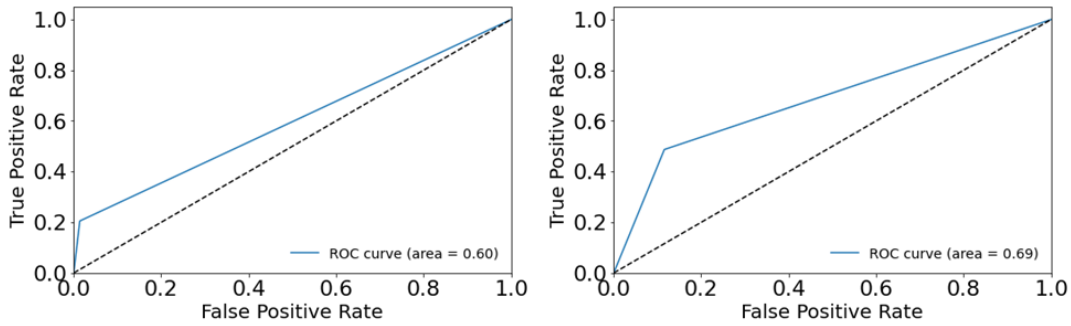


Figura 15: Differenti curve di ROC: Modello 1, Modello 4

5.3 Altri modelli decisionali

Come detto, sono stai poi applicati altri due algoritmi di classificazione. Il *KNN*, basato sull'osservazione dei valori simili più vicini, ha richiesto lo *scaling* delle varie *feature* del dataset. I valori migliori sono stati ottenuti impostando K pari a 11 per il *training set* originario e pari a 3 per quello espanso con lo *SMOTE*. Tuttavia, sono risultati in linea con quelli degli alberi decisionali.

Risultati migliori sono stati invece generati dall'algoritmo *Random Forest*, impostato per generare 100 diversi alberi. Come per gli alberi decisionali, le combi-

K	**Accuratezza train**	**Accuratezza con SMOTE**
1	0.742	0.757
2	0.809	0.752
3	0.786	0.731
4	0.822	0.731
5	0.802	0.713
6	0.824	0.697
7	0.812	0.692
8	0.828	0.688
9	0.824	0.673
10	0.831	0.684
11	0.829	0.678

Tabella 11: Accuratezze di diversi KNN con diversi valori di K e *training set*

nazioni migliori dei parametri sono stati impostate con l'ausilio del *random search*.

Nella tabella 12 vengono presentati i risultati ottenuti e una comparazione con due dei metodi ottenuti con gli alberi decisionali.

	Modello Random Forest 1	**Modello Random Forest 2**
Sampling	Nessuno	SMOTE
Criterion	Gini, Gini, Gini	Entropia, Gini, Entropia
Min Sample Split	5, 5, 5	5, 10, 20
Min Sample Leaf	1, 1, 5	1, 1, 1
Max Depth	26, 14, 8	27, 32, 19
Train Accuracy	0.974	0.987
Train F1-Score	1.90	1.97
Train AUC-ROC	0.922	0.987
Validation Accuracy	0.851	0.848
Validation F1-Score	1.24	1.29
Validation AUC-ROC	0.596	0.617

Tabella 12: Modelli *Random forest* (con i primi tre alberi) addestrati sul *training set* e sul *training set* su cui è stato applicato lo SMOTE

	Modello 1	**Modello 4**	**Modello 1 Random Forest**	**Modello 2 Random Forest (SMOTE)**
Training accuracy	0,851	0,931	0,973	0,987
Validation accuracy	0,844	0,763	0,851	0,847
Test Accuracy	0.86	0.82	0.96	0.97
Cross-validation = 10	0.83	0.85	0.85	0.90

Tabella 13: Confronto tra le accuratezze registrate con diversi modelli e diverse composizioni del *training set*

I risultati ottenuti sul *training set*, sia per il dataset originario che per quello espanso, sono sensibilmente migliori di quelli offerti dagli alberi decisionali. Si nota anche qui tuttavia una diminuzione con il *validation set*. Con il *test set*, invece, i due modelli si dimostrano particolarmente efficaci. In particolare il secondo, ottenuto con lo *SMOTE*, ha fornito risultati estremamente positivi, anche osservando la curva di ROC.

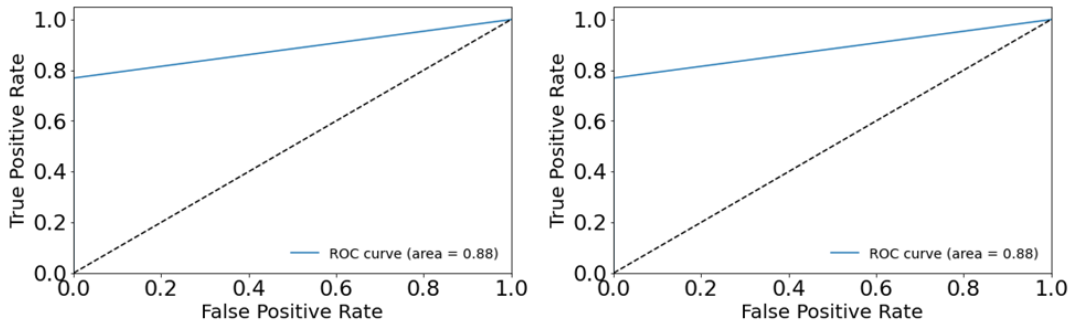


Figura 16: Differenti curve di ROC: Random Forest 1, Random Forest 2

Di seguito, il primo albero generato dal modello Random Forest 2.

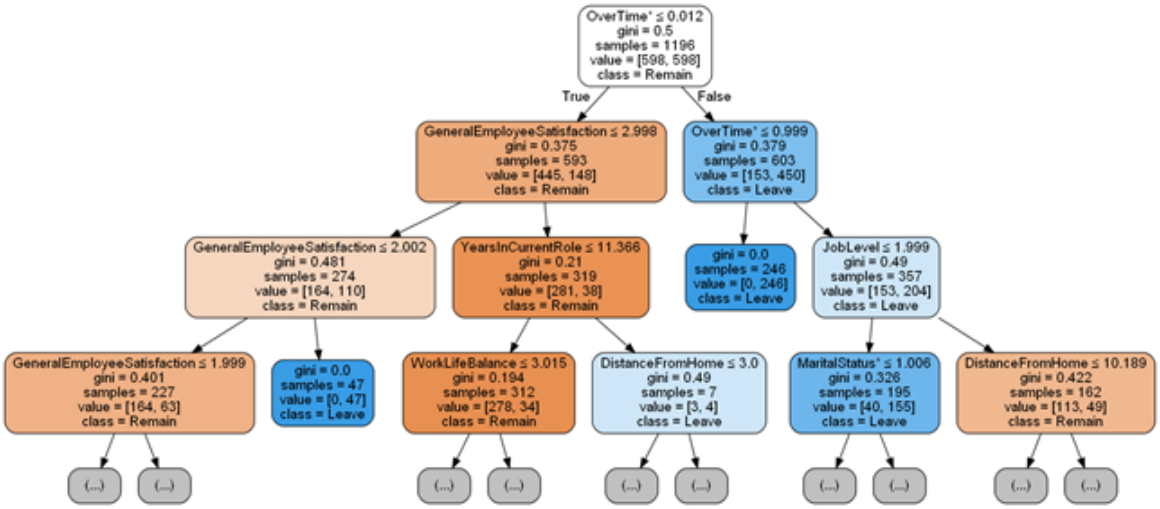


Figura 17: Albero generato dall'algoritmo *Random Forest*

Dall'analisi dell'albero presentato, e dai pesi assegnati da ciascuno dei classificatori addestrati alle diverse *feature*, si rileva l'importanza di quattro parametri particolarmente rilevanti per individuare l'*Attrition*: *OverTime*, *JobLevel*, *DistanceFromHome* e *GeneralEmployeeSatisfaction* (i primi tre da noi già utilizzati per il clustering). Meno rilevanti sono, ad esempio, *Gender* e *Age*.

Nel caso specifico sopra, si nota, ad esempio, che i lavoratori con basso *OverTime* tendono a rimanere in azienda. Maggiori valori di *OverTime* portano invece all'abbandono dell'azienda, con l'eccezione di alcuni casi determinati da *JobLevel* e *DistanceFromHome*.

Ulteriori analisi richiederebbero una lettura dell'albero a una maggiore profondità: tuttavia, è da ricordare che quello sopra proposto è soltanto uno dei 100 alberi del *Random Forest*.

5.4 Conclusioni sulla classificazione

In generale sono stati notati problemi legati alla natura estremamente sbilanciata del dataset, con una predominanza di oggetti con *Attrition* pari a *No*, che non hanno permesso ai classificatori di riconoscere correttamente i possibili lavoratori che abbandonano l'azienda.

Risultati migliori sono stati ottenuti sul *training set* con l'applicazione del metodo *SMOTE*. Tali modelli non sono comunque risultati adeguati per il *test set* (e il *validation set*), soffrendo di *overfitting*. Il problema si è presentato anche con il classificatore *KNN*, mentre il *Random Forest* ha costituito una valida soluzione, in quanto è un algoritmo nato con lo scopo di evitare le problematiche di *overfitting* che affliggono i tradizionali classificatori basati su alberi di decisione.

6 Association Rules

Questa sezione è dedicata all'*Association Rules Mining*: dopo una prima fase di *data preparation* abbiamo estratto gli *itemset* più frequenti e da questi, attraverso la funzione *Apriori*, ottenuto le regole di associazione. Il nostro obiettivo era quello di trovare delle regole con le quali sostituire i *missing values* presenti nel dataset e costruire un modello predittivo di *Attrition*, la nostra *target variable*.

6.1 Preparazione dei dati

Per l'esecuzione delle regole di associazione i dati sono stati preparati come discusso nella sezione di *Data Understanding*, senza però sostituire i *missing values*. Gli attributi quantitativi (*Age* e *MonthlyIncome*) sono stati poi discretizzati nei seguenti intervalli stabiliti osservando i grafici generati con il metodo *KDE* (Fig.18): per *Age* (18.0, 40.0], (40.0, 48.0], (48.0, 62.0], mentre per *MonthlyIncome* (1009.0, 4000.0], (4000.0, 8000.0], (8000.0, 11416.0].

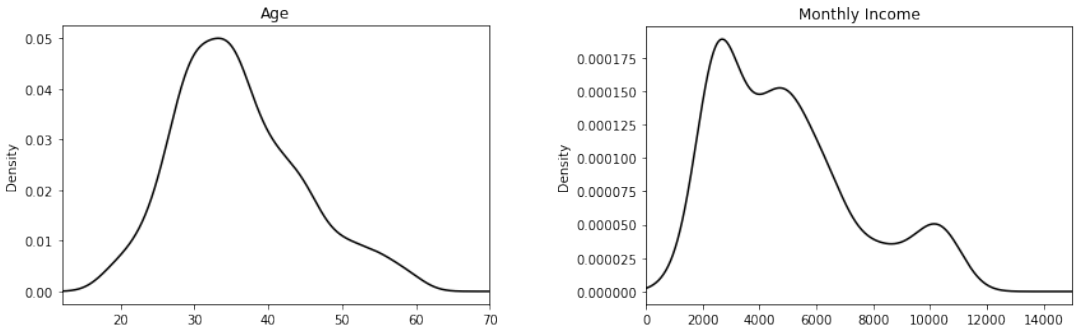


Figura 18: Kernel Density Estimation di Age e MonthlyIncome

6.2 *Frequent Itemset*

Ai dati preparati abbiamo applicato l'algoritmo *Apriori*, con *support* 10% e lunghezza minima dell'*itemset* 1, per ottenere i *frequent*, i *closed* e i *maximal itemset*. In totale abbiamo ottenuto 431 *frequent* e *closed itemset* e 148 *maximal itemset*. Nelle tabelle Tab. 14 e Tab. 15 abbiamo riportato solo gli itemset con alti valori di *support* (i *frequent itemset* e i *closed itemset* sono identici, e sono stati quindi inseriti nella stessa tabella). L'analisi dei *frequent itemset* evidenzia nuovamente, come già descritto nel *Data Understanding*, che la maggior parte dei lavoratori IBM sono uomini con età compresa tra 18 e 40 anni che possiedono un alto livello di soddisfazione lavorativa e personale e che, dunque, tendono a non licenziarsi.

Support	**Frequent-Closed itemsets**
80% - 90%	**1)** {Attrition: No} (*supp* = 0.83)
70% - 80%	**2)** {OverTime: No} (*supp* = 0.70)
60% - 70%	**3)** {OverTime: No, Attrition: No} (*supp* = 0.63) **4)** {Age: (18.0, 40.0]} (*supp* = 0.61) **5)** {WorkLifeBalance: High} (*supp* = 0.60)
50% - 60%	**6)** {Gender: Male} (*supp* = 0.57) **7)** {GeneralEmployeeSatisfaction: High} (*supp* = 0.53) **8)** {WorkLifeBalance: High, Attrition: No} (*supp* = 0.51) **9)** {Age: (18.0, 40.0], Attrition: No} (*supp* = 0.50)

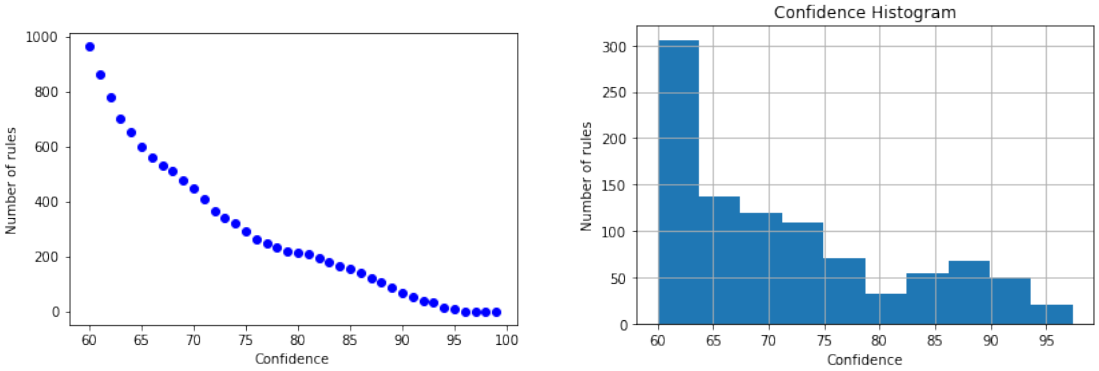
Tabella 14: *Frequent e Closed itemset con un valore di support > 50%*

Support	Maximal itemsets
15%	**1)** {JobLevel: Low, WorkLifeBalance: High, OverTime: No, Attrition: No} **2)** {JobLevel: Low, Age: (18.0, 40.0], OverTime: No, Attrition: No}
14%	**3)** {JobLevel: Very Low, WorkLifeBalance: High, OverTime: No, Attrition: No} **4)** {Gender: Male, WorkLifeBalance: High, Age: (18.0, 40.0], OverTime: No, Attrition: No} **5)** {MonthlyIncome: (1009.0, 4000.0], Age: (18.0, 40.0], OverTime: No, Attrition: No} **6)** {JobLevel: Low, Gender: Male, Overtime: No, Attrition: No} **7)** {Gender: Female, WorkLifeBalance: High, OverTime: No, Attrition: No} **8)** {WorkLifeBalance: Medium, Overtime: No, Attrition: No} **9)** {JobLevel: Very Low, Age: (18.0, 40.0], OverTime: No, Attrition: No}

Tabella 15: *Maximal itemsets ottenuti con min_support = 0.10*

6.3 Association Rules

Dai *frequent itemset* ottenuti abbiamo estratto le regole di associazione impostando una *min_confidence* del 60% (regole con *confidence* più bassa, vere cioè meno del 60% delle volte, sono poco informative). Abbiamo così ottenuto 965 regole. Alcune di esse, come si può notare dai grafici seguenti (Fig. 19), presentano una *confidence* compresa tra 80% e 90%.



(a) **(b)**

Figura 19: Distribuzione del numero di regole in base ai valori di *confidence*

Per quanto riguarda il *lift* (Fig. 20), la sua distribuzione è sbilanciata verso il valore di 1: vi sono, quindi, molte regole i cui elementi sono associati tra loro in maniera casuale. Solo un piccolissimo numero di regole ha lift compreso tra 1.6 e 1.9.

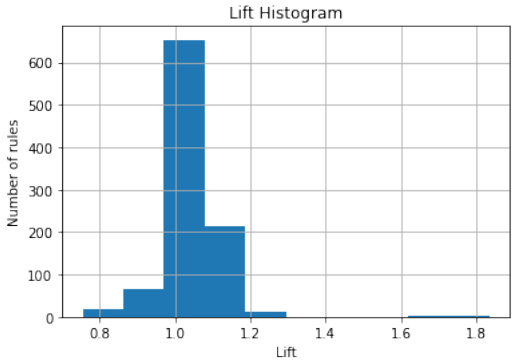


Figura 20: Distribuzione del numero di regole in base ai valori di lift

Delle 965 regole abbiamo selezionato le più informative. Per il principio di anti-monotonicità su cui si basa l'algoritmo *Apriori*, riportiamo successivamente solo 8 regole, ordinate secondo il parametro di lift, generate da diversi itemset.

Association Rules
{Attrition: Yes, Age: (18.0, 40.0]} => JobLevel: Very Low (*conf* = 0.71) (*lift* = 1.83)
{MonthlyIncome: (1009.0, 4000.0], Marital Status: Married} => Age: (18.0, 40.0] (*conf* = 0.74) (*lift* = 1.21)
{JobLevel: Very Low, Gender: Male, WorkLifeBalance: High, Attrition: No} => OverTime: No (*conf* = 0.86) (*lift* = 1.21)
{MonthlyIncome: (1009.0, 4000.0], JobLevel: Very Low} => Age: (18.0, 40.0] (*conf* = 0.74) (*lift* = 1.21)
{OverTime: Yes, Gender: Male, Attrition: No} => GeneralEmployeeSatisfaction: High (*conf* = 1.83) (*lift* = 0.64)
{JobLevel: Low, GeneralEmployeeSatisfaction: High, Age: (18.0, 40.0]} => Attrition: No (*conf* = 0.97) (*lift* = 1.17)
{Gender: Male, WorkLifeBalance: High, Age: (18.0, 40.0], OverTime: No} => Attrition: No (*conf* = 0.95) (*lift* = 1.15)
{GeneralEmployeeSatisfaction: Medium, MonthlyIncome: (1009.0, 4000.0]} => Gender: Male (*conf* = 0.64) (*lift* = 1.12)

Tabella 16: *Association Rules più interessanti, ordinate per valore di lift*

Osservando queste regole possiamo affermare che circa il 96% dei lavoratori IBM non si licenzia e nell'86% dei casi non fa straordinari. Con probabilità del 75%, i dipendenti hanno età compresa tra 18 e 40 anni e sono uomini nel 64% dei casi. La quinta regola, infine, evidenzia che nel 64% dei casi i lavoratori sono molto soddisfatti del loro lavoro (dimostrando che la nuova *feature* da noi creata (*GeneralEmployeeSatisfaction*) è risultata utile nelle analisi).

6.4 Sostituzione dei *missing values*

Dopo aver estratto le regole più interessanti, abbiamo sostituito i valori mancanti di *Age*, *MonthlyIncome* e *Gender*.

- *Age* presentava 168 *missing values* e con le regole estratte (con *confidence* compresa tra 60% e 75%) siamo riusciti a sostituirne 89 con il valore (18.0, 40.0]. Gli altri intervalli d'età presentavano una *confidence* minore nel 60% e non sono stati considerati validi per la sostituzione;
- *MonthlyIncome* aveva 213 *missing values* che non siamo riusciti a sostituire in quanto le prime regole utili per la sostituzione hanno tutte una *confidence* del 40%;
- *Gender* aveva 51 *missing values* e con le regole estratte (con *confidence* compresa tra 60% e 64%) siamo riusciti a sostituirne 40 con il valore *Male*. Per il valore di *Female* non abbiamo trovato regole con una *confidence* maggiore del 60%.

6.5 Predizione della *target variable*

Abbiamo usato le regole estratte per costruire un modello predittivo per la nostra *target variable*. L'accuratezza del modello è del 94%. Sebbene possa apparire un ottimo risultato, il modello assegna sempre e solo il valore di *Attrition: No* dal momento che non vi sono regole con *confidence* maggiore del 60% che hanno come conseguenza *Attrition: Yes*.

Tuttavia abbiamo deciso di rilanciare il modello includendo la prima regola che comprendesse *Attrition: Yes*, che presenta *confidence* pari a 57%, un valore di poco minore di 60%. La nuova accuratezza è risultata dell'82%, ancora un valore alto, ma a causa del forte sbilanciamento dei dati il modello ha continuato ad associare solo il valore di *Attrition: No*.

7 Conclusioni

Riassumendo, l'analisi del dataset si è ripetutamente scontrata con il suo alto sbilanciamento e la presenza di dati mancanti o corrotti, già rilevati nella fase di *Data understanding*.

Nella fase di *Data preparation* si è cercato limitare il problema, rimuovendo i parametri giudicati poco utili, stimando i possibili valori mancanti nelle varie *feature* e generando due nuove *feature ad hoc*, che si sono in seguito rivelate utili. Nel complesso, i 1176 oggetti descritti da 33 attributi sono stati ridotti a 1029 oggetti descritti da 13 attributi.

Il clustering dei dati è risultato tuttavia comunque problematico per il generale sbilanciamento degli oggetti con *Attrition* pari *No* e la presenza di rumore. Dei vari modelli testati, solo il *K-Means* ha portato a dei buoni risultati.

Successivamente si è proceduto alla classificazione dei dati. Nuovamente, le problematiche descritte sopra hanno portato gli algoritmi di *Decision tree* e *KNN* a generare risultati poco soddisfacenti. Di contro, il metodo *Random forest*, accompagnato da un *oversampling* del dataset, è risultato efficace e ha permesso di comprendere quali attributi potessero maggiormente determinare l'*Attrition*: *OverTime*, *JobLevel*, *GeneralEmployeeSatisfaction* e *DistanceFromHome*. È stato inoltre addestrato un modello per la predizione della *target variable*, che, tuttavia, non ha fornito risultati accurati.

Eventuali sviluppi futuri potrebbero vedere l'applicazione di altri algoritmi o in alternativa l'ampliamento dell'attuale dataset, per avere una prospettiva meno sbilanciata dei lavoratori IBM, permettendo, anche soltanto con le metodologie testate in questo *paper*, di ottenere risultati più accurati.