Università della Calabria

Dipartimento di Economia, Statistica e Finanza "*Giovanni Anania*"

Corso di Laurea Magistrale in Statistica e Informatica per le Decisioni e le Analisi di Mercato

Processo di scoperta della conoscenza, tramite di tecniche di apprendimento supervisionato e non supervisionato, per rafforzare l'Employee Retention

Docenti

Prof. Andrea Tagarelli
Prof. Eugenio Vocaturo

Studenti

Tucci Marco Matr. 214216
Ruffo Pietro Matr. 214241

Anno Accademico 2020/2021

INDICE

Introduzione

Il termine *retention* indica in generale un'azione di "mantenimento" o "conservazione". In ambito aziendale questo aspetto può assumere due facce: da una parte la fidelizzazione del cliente, dall'altra le strategie di conservazione dei propri dipendenti. Questa seconda forma di *retention* è detta Employee Retention o HR Retention e rappresenta una delle principali problematiche della HR Analytics.
Spesso è un aspetto sottovalutato, ma le persone che lavorano all'interno di un'impresa fanno davvero la differenza e costituiscono un punto di forza per la stessa.

Se all'interno del team ci sono equilibrio e collaborazione, la defezione di un componente può rappresentare un grosso problema, oltre che uno spreco di risorse se si considera quanto è costata la formazione del lavoratore.

Per tanto si rivela di fondamentale importanza l'Employee Retention, ovvero la capacità dell'azienda di attuare strategie e politiche per attrarre dipendenti talentuosi e conservarli a lungo.

Un team di lavoro composto da dipendenti competenti, motivati e soddisfatti è il primo passo verso il successo dell'azienda.

Obiettivo dell'analisi è individuare i fattori chiave che inducono il lavoratore dipendente a rassegnare le proprie dimissioni così da poter mettere in atto strategie preventive atte non solo a prevenire eventuali defezioni ma anche ad individuare quei fattori volti ad aumentare la fidelizzazione e il senso di appartenenza del dipendente verso l'azienda.

Lo studio verrà condotto impiegando tecniche di apprendimento supervisionato e non supervisionato implementate nel tool "WEKA".

Attraverso l'impiego di algoritmi di Machine Learning sarà possibile non solo definire clusters di soggetti con caratteristiche simili tra loro, ma anche ottenere il miglior modello di classificazione per individuare i dipendenti che presenteranno le proprie dimissioni e quelli che invece resteranno saldamente nell'organico aziendale.

L'analisi dei dati a disposizione verrà effettuata facendo riferimento al processo di scoperta della conoscenza (o di Knowledge Discovery in Databases - KDD) che prevede le seguenti fasi: estrazione e selezione dei dati, pre-processamento e trasformazione dei dati, data mining (in particolare, classificazione e clustering), interpretazione e validazione dei risultati ottenuti.

1. Descrizione del Dataset

Il dataset impiegato per l'applicazione delle tecniche di Data Mining è "IBM HR Analytics Employee Attrition & Performance", un set di dati fittizio contenente informazioni sul personale dipendente dell'omonima azienda informatica per un totale di 1470 tuple definite su 35 attributi:

Variabile	Tipo	Descrizione
Age	Discreta	Età del dipendente in anni
Attrition	Binaria	Indica se il dipendente ha rassegnato le sue dimissioni o meno
BusinessTravel	Categoriale	Indica quanto spesso il dipendente viaggia per raggiungere il posto di lavoro.
DailyRate	Discreta	Indica il costo giornaliero del singolo lavoratore per l'azienda
Department	Categoriale	Divisione aziendale di appartenenza del singolo dipendente
DistanceFromHome	Discreta	Distanza in miglia tra il posto di lavoro e l'abitazione del dipendente
Education	Discreta	Tipologia di formazione del singolo dipendente
EducationField	Categoriale	Area di specializzazione relativa all'istruzione del dipendente
EmployeeCount	Discreta	Misura di supporto presente nel dataset
EmployeeNumber	Discreta	Codice identificativo del dipendente
EnvironmentSatisfaction	Discreta	Indica il grado di soddisfazione del dipendente rispetto all'ambiente aziendale espresso su una scala da 1 a
Gender	Categoriale	Indica il sesso del dipendente
HourlyRate	Discreta	Tariffa oraria del singolo dipendente. Indica quanto viene pagata l'ora lavorata
JobInvolvement	Discreta	Livello di coinvolgimento relativo al singolo dipendente in azienda espresso su una scala da 1 a 4
JobLevel	Discreta	Livello relativo alla posizione ricoperta in azienda dal

		singolo dipendente espresso su una scala da 1 a 5
JobRole	Categoriale	Posizione ricoperta in azienda da parte del dipendente
JobSatisfaction	Discreta	Livello di soddisfazione del proprio lavoro espresso su una scala da 1 a 4
MaritalStatus	Categoriale	Stato civile del dipendente
MonthlyIncome	Discreta	Stipendio mensile del lavoratore
MonthlyRate	Discreta	Indica il costo mensile del singolo lavoratore per l'azienda
NumCompaniesWorked	Discreta	Numero di aziende per le quali ha lavorato il singolo dipendente
Over18	Categoriale	Indica se il dipendente ha compiuto 18 anni di età
OverTime	Categoriale	Indica se il dipendente ha effettuato delle ore di lavoro straordinario
PercentSalaryHike	Discreta	Percentuale di aumento della retribuzione dal primo contratto di lavoro relativo al dipendente
PerformanceRating	Discreta	Livello di performance relative al singolo dipendente. Assume due valori: 3(sufficiente) o 4(eccellente)
RelationshipSatisfaction	Discreta	Grado di soddisfazione, su una scala da 1 a 4, delle relazioni in ambito lavorativo relativo al singolo dipendente
StandardHours	Discreta	Ore settimanali di lavoro relative al dipendente
StockOptionLevel	Discreta	Titolo derivato che dà il diritto di acquistare azioni della società ad un determinato prezzo definito dal livello del titolo stesso. Ci sono 4 livelli: 0, 1, 2 e 3.
TotalWorkingYears	Discreta	Numero totale di anni di lavoro del dipendente
TrainingTimesLastYear	Discreta	Numero training di formazione ricevuti l'anno precedente dal lavoratore
WorkLifeBalance	Discreta	Equilibrio vita professione e personale del dipendente espresso su una scala

		discreta da 1 a 4
YearsAtCompany	Discreta	Numero anni di lavoro in azienda del dipendente
YearsInCurrentRole	Discreta	Numero anni nell'attuale posizione per il dipendente
YearsSinceLastPromotion	Discreta	Numero anni dall'ultima promozione ricevuta dal dipendente
YearsWithCurrManager	Discreta	Numero di anni di lavoro con l'attuale manager per il dipendente

Figura 1.1 Tabella di descrizione degli attributi

2. Fase di Pre-Processing e Trasformazione

Questa fase consiste nel pulire i dati, rimuovendo "rumori" o altre inconsistenze che potrebbero causare problemi al processo di analisi. Inoltre, provvede ad esplorare e preparare i dati per gli step successivi atti all'individuazione dei pattern.

Per prima cosa, utilizzando la funzione "Remove" si eliminano gli attributi "EmployNumber", "EmployCount" e "StandardHours". L'attributo "EmployNumber" funge da mero codice identificativo, "EmployCount" è una misura di supporto e l'attributo "StandardHours" presenta valori uguali per tutte le tuple. Per tanto, tutti e tre gli attributi citati non sono utili ai fini dell'analisi.

Trattamento dei dati mancanti

Successivamente viene individuato un missing value nell'attributo "Department" che verrà sostituito con la moda tramite la funzione "ReplaceMissingValues:

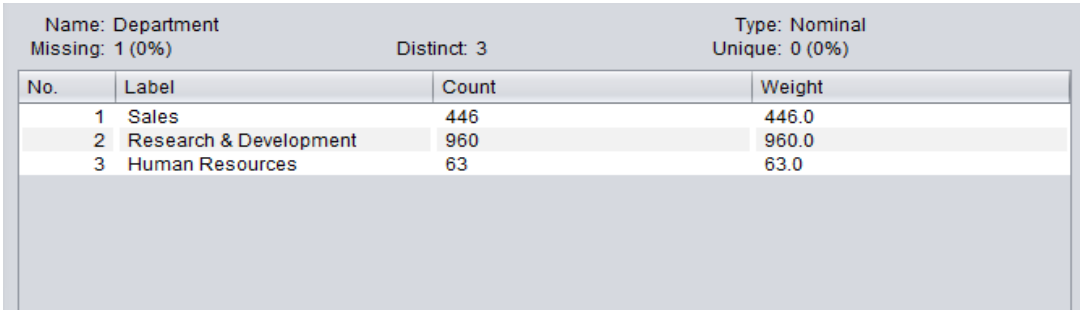


Figura 2.1 Individuazione missing value nell'attributo "Department"

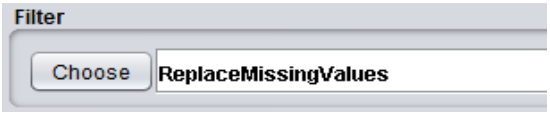


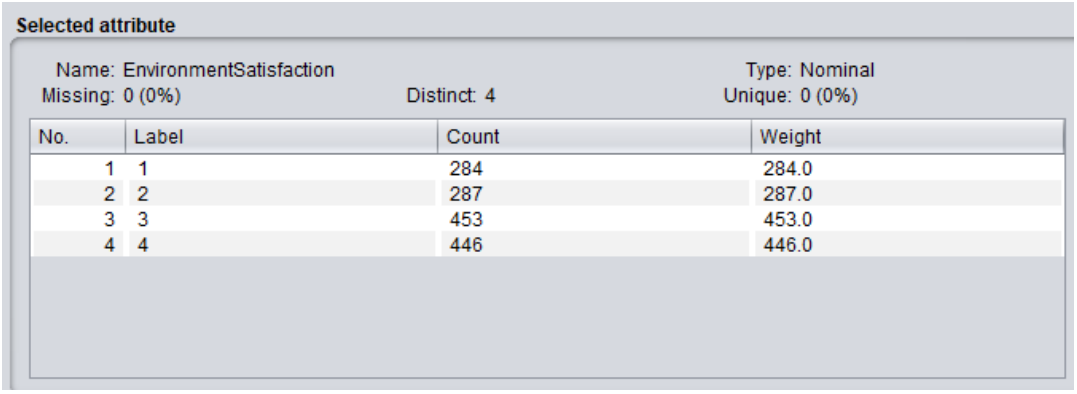
Figura 2.2 Sostituzione missing value presente in "Deparment" con la moda

Operazioni di trasformazione delle variabili

Dopodiché, le seguenti variabili verranno trasformate in variabili nominali:

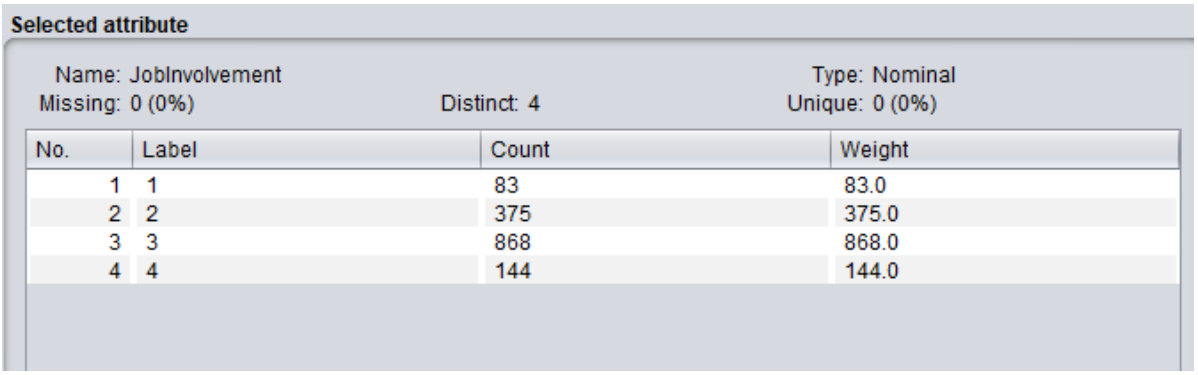
- "EnvironmentSatisfaction";
- "JobInvolvment";
- "JobLevel";
- "StockOptionLevel";

- “Education”;
- “JobSatsfaction”;
- “WorkLifeBalance”;
- “RelationshipSatisfaction”.



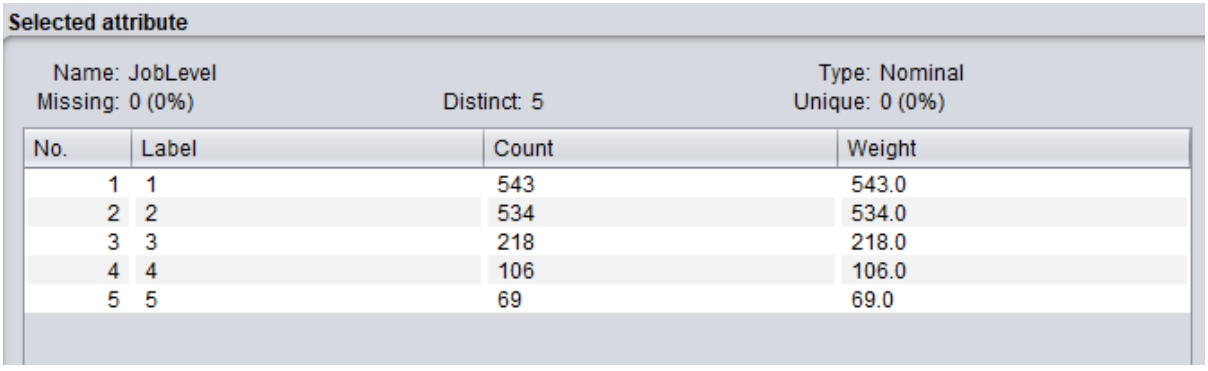
No.	Label	Count	Weight
1	1	284	284.0
2	2	287	287.0
3	3	453	453.0
4	4	446	446.0

Figura 2.3 Trasformazione attributo “EnvironmentSatisfaction” in variabile nominale



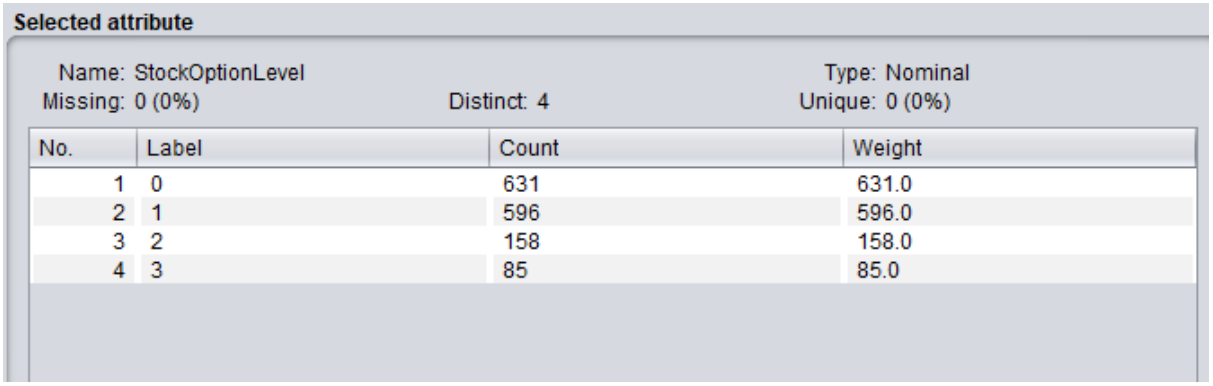
No.	Label	Count	Weight
1	1	83	83.0
2	2	375	375.0
3	3	868	868.0
4	4	144	144.0

Figura 2.4 Trasformazione attributo “JobInvolvement” in variabile nominale



No.	Label	Count	Weight
1	1	543	543.0
2	2	534	534.0
3	3	218	218.0
4	4	106	106.0
5	5	69	69.0

Figura 2.5 Trasformazione attributo “JobLevel” in variabile nominale

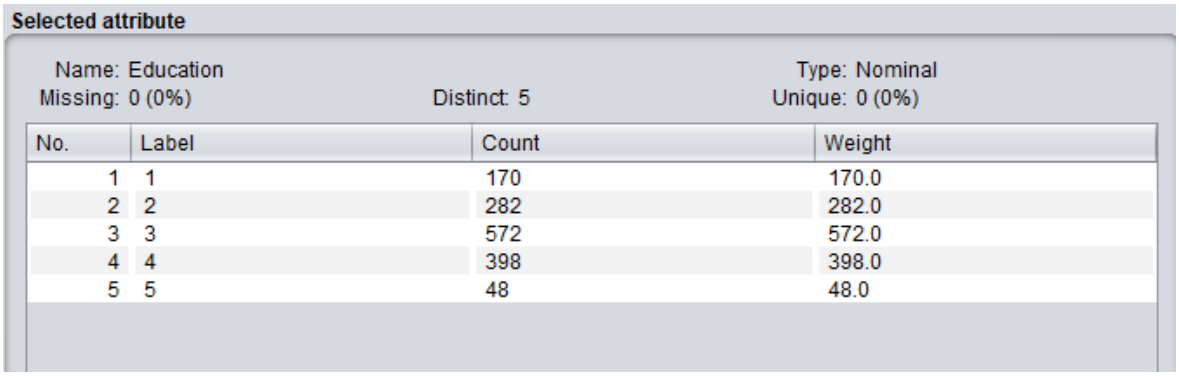


Selected attribute

Name: StockOptionLevel — Type: Nominal
Missing: 0 (0%) — Distinct: 4 — Unique: 0 (0%)

No.	Label	Count	Weight
1	0	631	631.0
2	1	596	596.0
3	2	158	158.0
4	3	85	85.0

Figura 2.6 Trasformazione attributo "StockOptionLevel" in variabile nominale

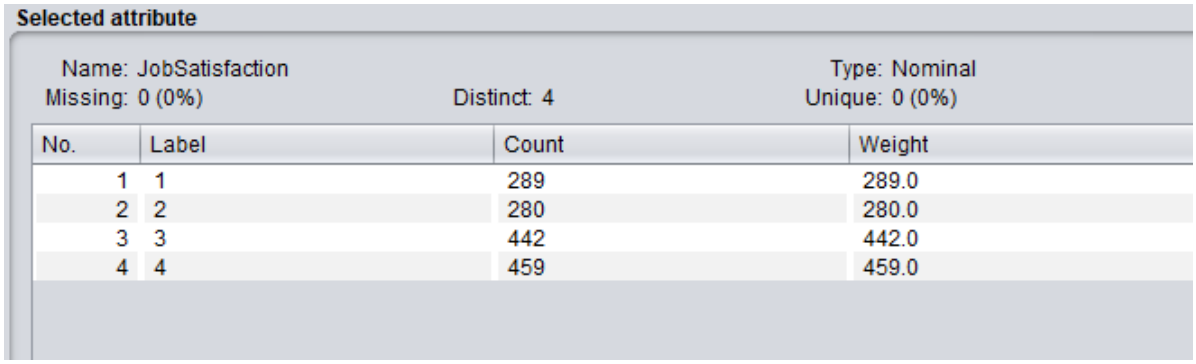


Selected attribute

Name: Education — Type: Nominal
Missing: 0 (0%) — Distinct: 5 — Unique: 0 (0%)

No.	Label	Count	Weight
1	1	170	170.0
2	2	282	282.0
3	3	572	572.0
4	4	398	398.0
5	5	48	48.0

Figura 2.7 Trasformazione attributo "Education" in variabile nominale

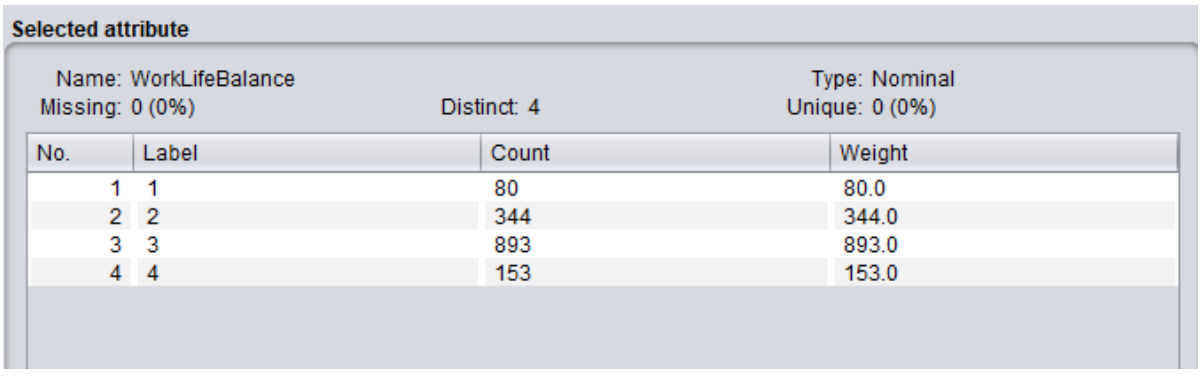


Selected attribute

Name: JobSatisfaction — Type: Nominal
Missing: 0 (0%) — Distinct: 4 — Unique: 0 (0%)

No.	Label	Count	Weight
1	1	289	289.0
2	2	280	280.0
3	3	442	442.0
4	4	459	459.0

Figura 2.8 Trasformazione attributo "JobSatisfaction" in variabile nominale



Selected attribute

Name: WorkLifeBalance — Type: Nominal
Missing: 0 (0%) — Distinct: 4 — Unique: 0 (0%)

No.	Label	Count	Weight
1	1	80	80.0
2	2	344	344.0
3	3	893	893.0
4	4	153	153.0

Figura 2.9 Trasformazione attributo "WorkLifeBalance" in variabile nominale

Selected attribute

Name: RelationshipSatisfaction Type: Nominal
Missing: 0 (0%) Distinct: 4 Unique: 0 (0%)

No.	Label	Count	Weight
1	1	276	276.0
2	2	303	303.0
3	3	459	459.0
4	4	432	432.0

Figura 2.10 Trasformazione attributo "RelationshipSatisfaction" in variabile nominale

Importanza degli attributi

Al fine di individuare gli attributi maggiormente rilevanti si applica il filtro "AttributeSelection".

Si tratta di un filtro di attributi supervisionato che può essere usato per selezionare gli attributi. È molto flessibile e permette di combinare vari metodi di ricerca e valutazione. Valuta il valore di un sottoinsieme di attributi considerando la capacità predittiva individuale di ogni caratteristica e il grado di ridondanza tra le stesse. Selezionando come attributo di classe "Attrition", si ottiene il seguente risultato.

No.	Name
1	Age
2	BusinessTravel
3	EnvironmentSatisfaction
4	JobInvolvement
5	JobSatisfaction
6	MonthlyIncome
7	OverTime
8	StockOptionLevel
9	TotalWorkingYears
10	WorkLifeBalance
11	YearsAtCompany
12	YearsWithCurrManager
13	Attrition

Figura 2.11 Attributi risultanti dall'applicazione del filtro "AttributeSelection"

Per comprendere meglio l'importanza delle variabili di input rispetto all'attributo di classe, è possibile analizzare l'impatto che tali variabili hanno sulle dimissioni del dipendente utilizzando l'indicatore "GainRatio" calcolato su ogni attributo precedentemente individuato.

Il grafico successivo mostra che la variabile "StockOptionLevel" presenta il massimo valore di "GainRatio".

Si ipotizza che questo sia l'attributo per cui viene effettuato lo split del nodo radice nel caso in cui venga utilizzato un algoritmo di classificazione, garantendo la massima riduzione di entropia passando dal nodo radice ai nodi figli.

Attribute selection output

```
Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 13 Attrition):
        Gain Ratio feature evaluator

Ranked attributes:
 0.146     8 StockOptionLevel
 0.114    12 YearsWithCurrManager
 0.1098   11 YearsAtCompany
 0.1087    7 OverTime
 0.0881    1 Age
 0.0852    9 TotalWorkingYears
 0.058     6 MonthlyIncome
 0.0412    3 EnvironmentSatisfaction
 0.0354    2 BusinessTravel
 0.033     5 JobSatisfaction
 0.0283    4 JobInvolvement
 0.0131   10 WorkLifeBalance

Selected attributes: 8,12,11,7,1,9,6,3,2,5,4,10 : 12
```

Figura 2.12 Calcolo "GainRatio"

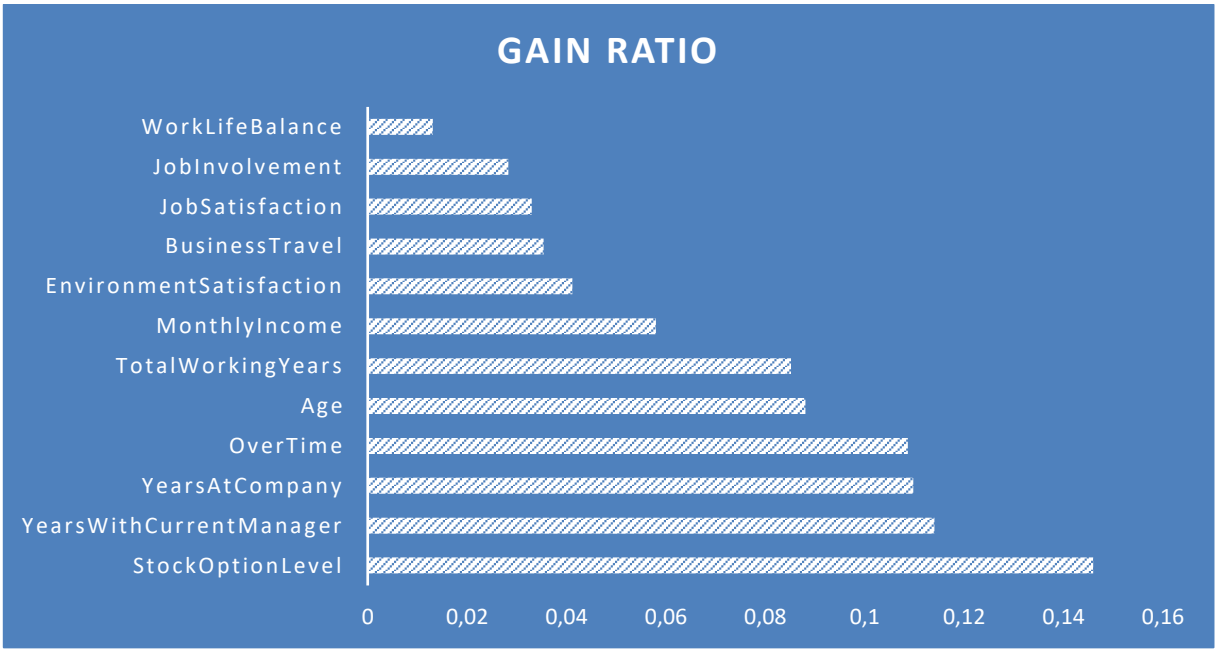


Figura 2.13 Grafico "GainRatio" per i vari attributi

Si verifica, inoltre, il legame tra tutte le variabili e l'attributo di classe "Attrition" tramite il calcolo del coefficiente di correlazione di Pearson.

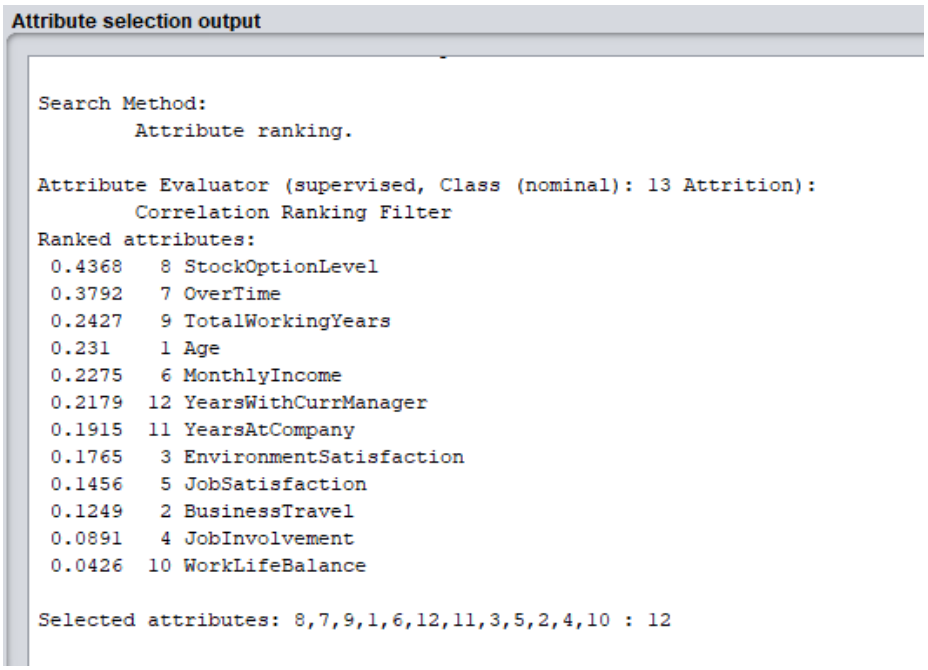


Figura 2.14 Correlazione tra l'attributo di classe "Attrition" e gli altri attributi.

Da queste analisi si evince come le variabili "StockOptionLevel" e "OverTime" siano quelle maggiormente correlate all'attributo "Attrition". Quindi la decisione di un dipendente circa l'abbondono dell'organico aziendale sembrerebbe essere influenzata dalla possibilità dello stesso di poter acquistare titoli dell'impresa e dalla sua disponibilità a fare straordinari.

Analisi delle distribuzioni

La figura che segue mostra le distribuzioni delle variabili selezionate considerando le due modalità dell'attributo di classe "Attrition".

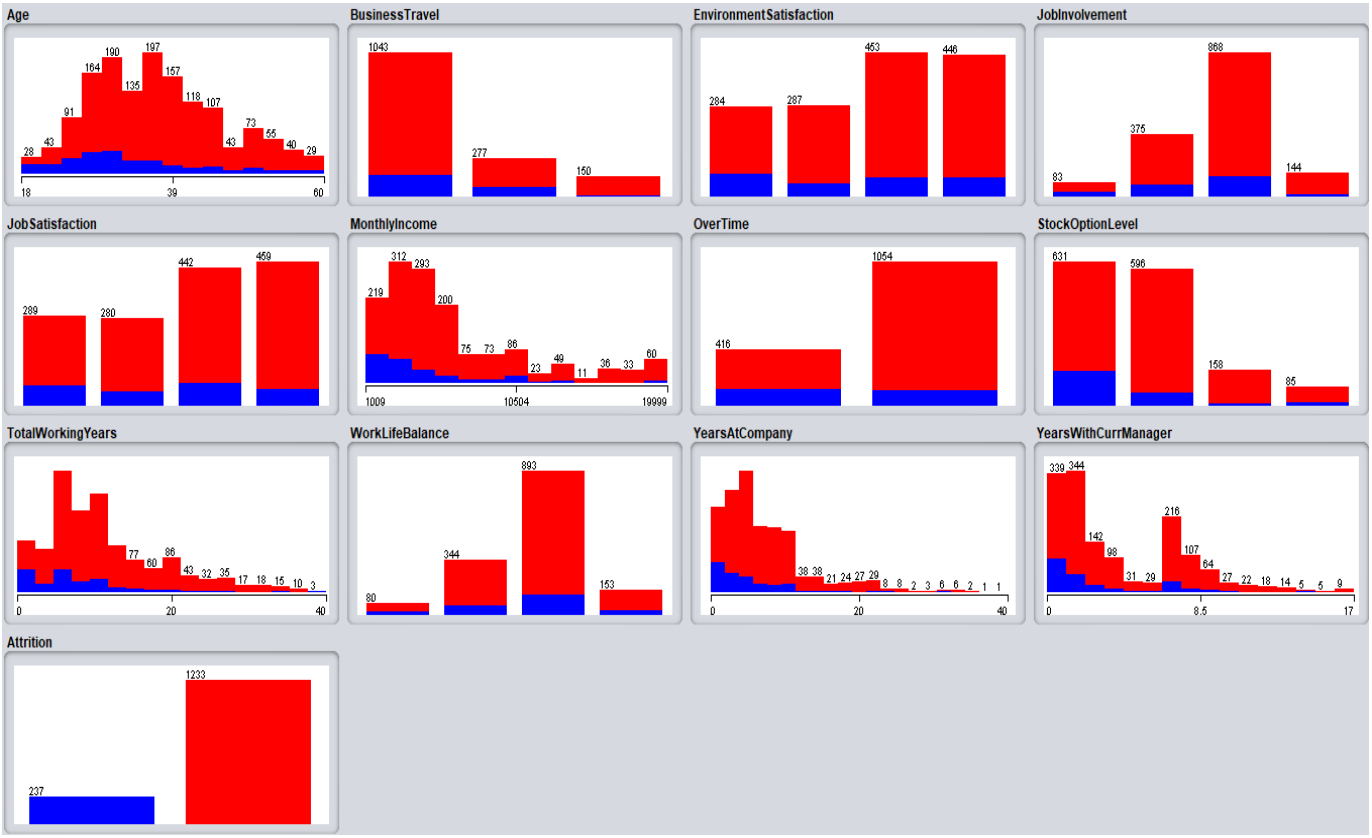


Figura 2.15 distribuzione attributo di classe "Attrition" rispetto gli altri attributi

Si nota immediatamente uno sbilanciamento nel dataset rispetto all'attributo di classe. Al fine di trattare questo problema, si usa la funzione "SMOTE" (Synthetic Minority Oversampling Technique) che permette di incrementare la classe sbilanciata. Piuttosto che replicare le osservazioni a bassa frequenza (nel caso specifico i dipendenti che hanno rassegnato le dimissioni), questa funzione permette di creare osservazioni sintetiche basate sulle osservazioni di minoranza esistenti.

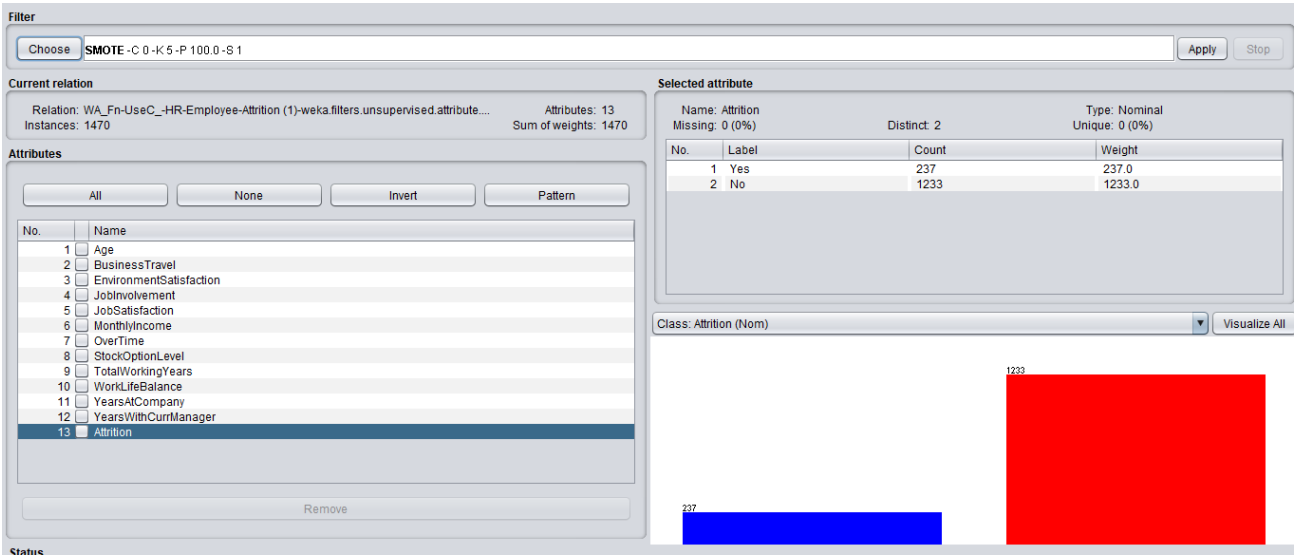


Figura 2.16 Distribuzione di frequenza dell'attributo di classe "Attrition" pre SMOTE

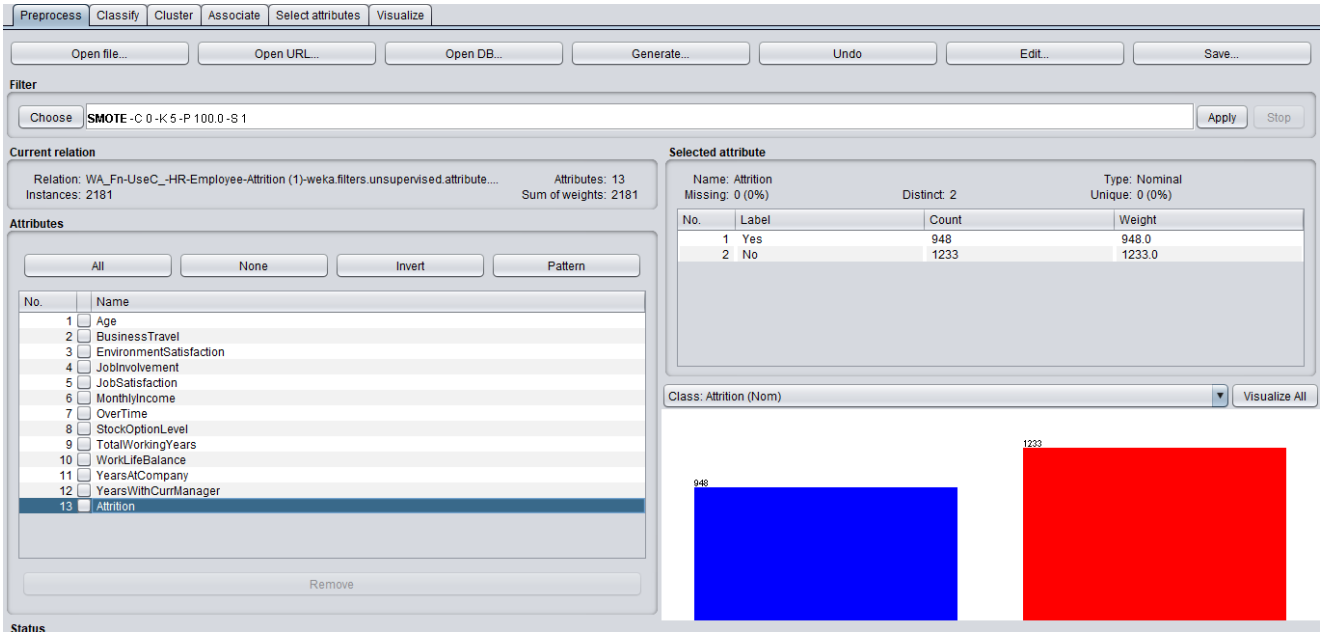


Figura 2.17 Distribuzione di frequenza dell'attributo di classe "Attrition" post SMOTE

Si è proceduto successivamente ad un posizionamento casuale delle istanze all'interno del dataset tramite un'appropriata funzione: il filtro non supervisionato "Randomize".

Normalizzazione

Al fine di evitare che le variabili numeriche con ordine di grandezza molto elevato inficino il risultato finale degli algoritmi, si procede con la loro normalizzazione.

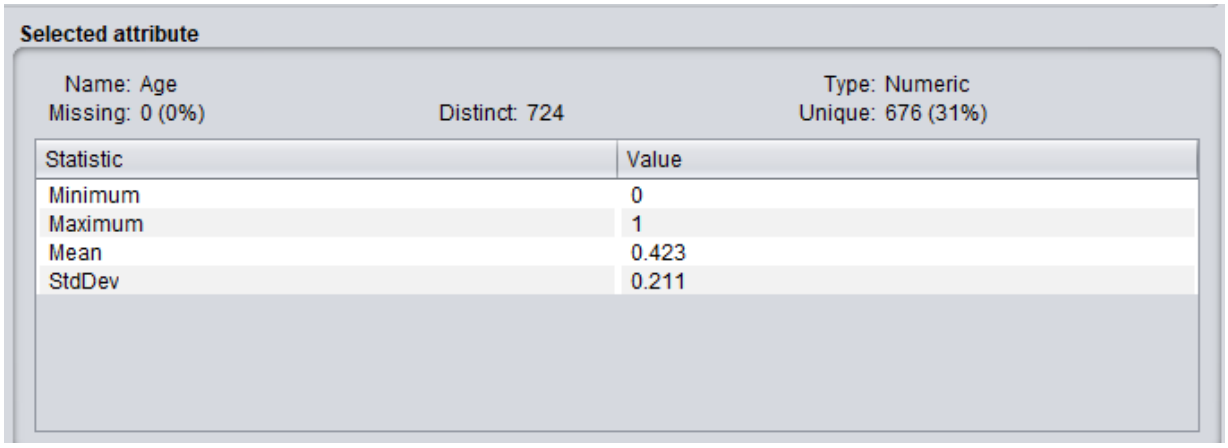


Figura 2.18 Normalizzazione attributo "Age"

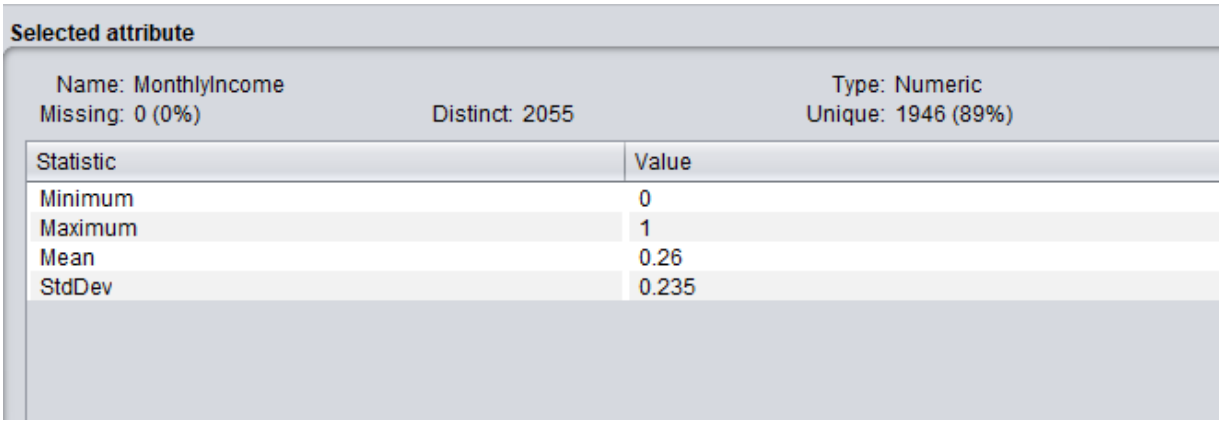


Figura 2.19 Normalizzazione attributo "MonthlyIncome"

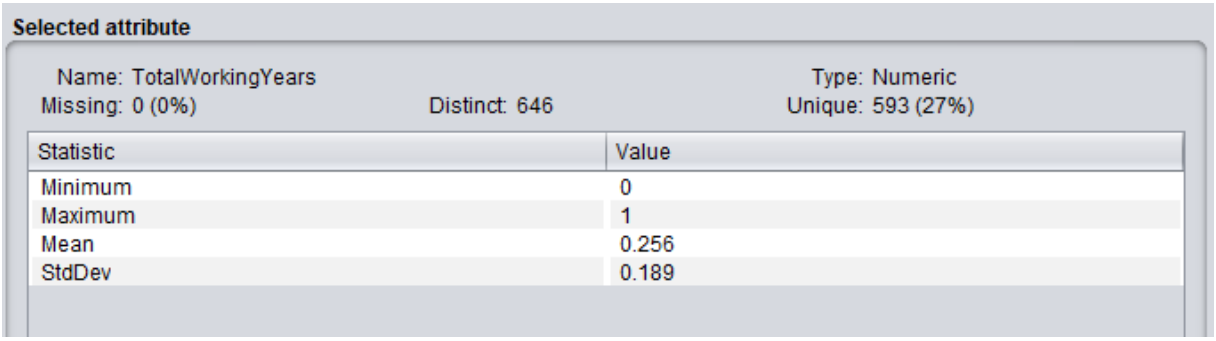


Figura 2.20 Normalizzazione attributo "TotalWorkingYears"

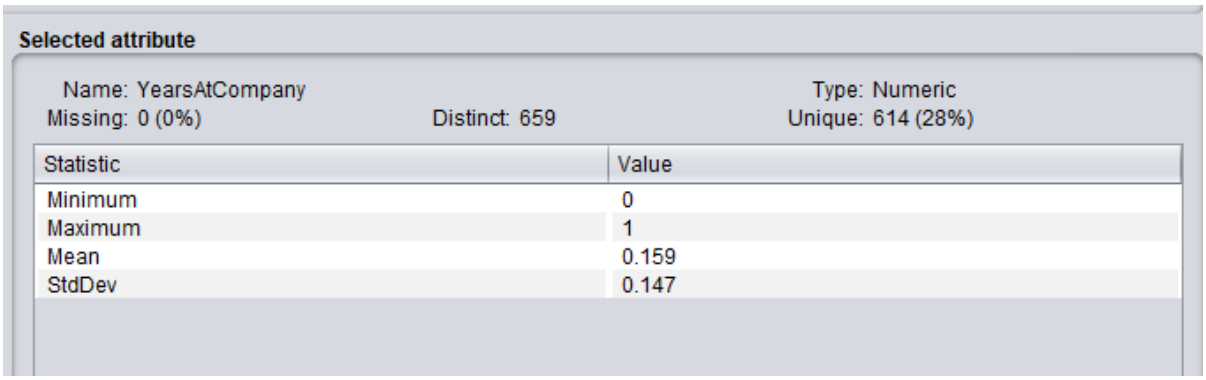


Figura 2.21 Normalizzazione attributo "YearsAtCompany"

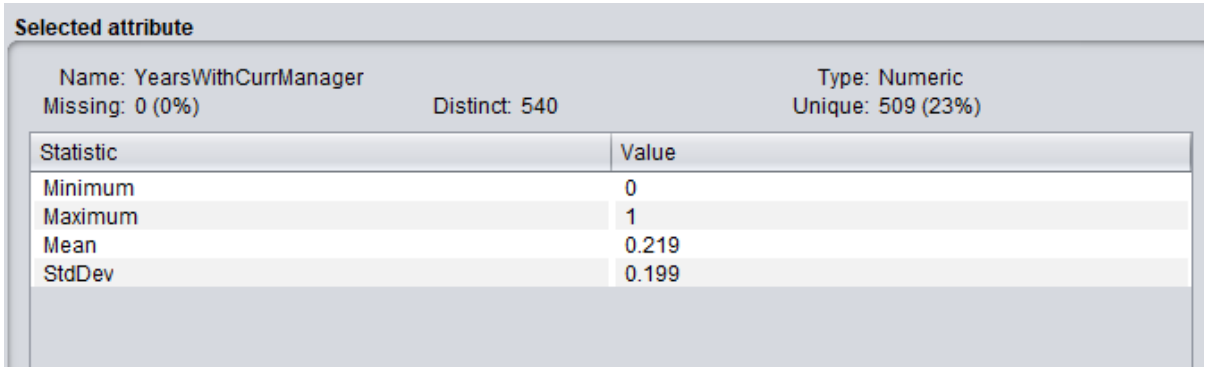


Figura 2.22 Normalizzazione attributo "YearsWithCurrManager"

3. Classificazione

Si passa quindi alla fase di processing. Il task di mining che verrà impiegato in questo capitolo è la classificazione (tecnica di apprendimento supervisionata). Saranno utilizzati diversi algoritmi che permetteranno di costruire classificatori in grado di prevedere il valore dell'attributo di classe ("Attrition" che indica se il dipendente ha dato le proprie dimissioni o meno) in funzione degli attributi selezionati nella fase di pre-processing. Si valuterà infine quale tra i classificatori, ottenuti dagli algoritmi considerati, meglio si presta a classificare le istanze presenti nel dataset. Dal momento che quest'ultimo non è molto numeroso (2181 tuple), si applicherà sempre la tecnica della "10-cross fold validation" per valutare i risultati ottenuti. L'obiettivo finale è quello di poter avvalersi di un modello matematico con proprietà di generalizzazione in grado di prevedere se un dipendente rassegnerà le proprie dimissioni o meno.

Seguono i risultati ottenuti su "WEKA" applicando i diversi algoritmi.

3.1 "NaiveBayes"

Con l'algoritmo "NaiveBayes" basato sul teorema di Bayes, e quindi su una rappresentazione probabilistica dei dati, si assume che tutti gli attributi in input abbiano la stessa importanza e siano indipendenti tra loro.

Di seguito si riportano i risultati derivanti dall'applicazione di questo algoritmo.

```
Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1738               79.6882 %
Incorrectly Classified Instances       443               20.3118 %
Kappa statistic                          0.5907
Mean absolute error                      0.2374
Root mean squared error                  0.3863
Relative absolute error                 48.303  %
Root relative squared error             77.919  %
Total Number of Instances             2181

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,809    0,212    0,745      0,809    0,776      0,592    0,873     0,875     Yes
               0,788    0,191    0,843      0,788    0,814      0,592    0,873     0,855     No
Weighted Avg.  0,797    0,200    0,801      0,797    0,798      0,592    0,873     0,864

=== Confusion Matrix ===

   a   b   <-- classified as
 767 181 |   a = Yes
 262 971 |   b = No
```

Figura 3.1 Output dell'algoritmo "NaiveBayes"

Il classificatore creato permette di ottenere un valore di accuratezza pari al 79,69%. Le istanze del dataset mal classificate sono 443. In particolare, attraverso la matrice di confusione, si può notare che 262 unità etichettate come "No" sono classificate come "Yes" e 181 unità etichettate come "Yes" sono state classificate come "No". Come è evidente, questo classificatore non ha prodotto risultati soddisfacenti, poiché sicuramente vi sarà qualche correlazione, seppur debole, tra qualche attributo. Inoltre, si evidenzia che il valore di "Kappa statistic", indicatore di robustezza del modello, non è elevato.

3.2 "AdaBoostM1"

"AdaBoost" è il primo algoritmo di "boosting" creato per la classificazione binaria. Rappresenta oggi una tecnica di boosting molto diffusa ed utilizzata in quanto combina più "classificatori deboli" in un solo "classificatore forte". Considerando più modelli, "AdaBoost" consente di ottenere un modello globalmente più performante rispetto ai singoli classificatori. Ad ogni iterazione, un nuovo "classificatore debole" viene valutato in sequenza cercando di creare un "classificatore forte", in modo da produrre stime più precise.

Segue l'output dell'algoritmo applicato.

```
Time taken to build model: 0.27 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1790               82.0724 %
Incorrectly Classified Instances       391               17.9276 %
Kappa statistic                          0.6369
Mean absolute error                      0.2615
Root mean squared error                  0.3537
Relative absolute error                 53.2126 %
Root relative squared error             71.3584 %
Total Number of Instances             2181

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,813    0,174    0,783      0,813    0,798      0,637    0,901     0,881     Yes
               0,826    0,187    0,852      0,826    0,839      0,637    0,901     0,917     No
Weighted Avg.  0,821    0,181    0,822      0,821    0,821      0,637    0,901     0,901

=== Confusion Matrix ===

   a    b   <-- classified as
 771  177 |    a = Yes
 214 1019 |    b = No
```

Figura 3.2 Output dell'algoritmo "AdaBoostM1"

In questo caso, l'accuratezza è più elevata e pari all'82,07%, infatti le istanze classificate in maniera non corretta sono 214. Si tratta di un classificatore leggermente più robusto del precedente.

3.3 "J48"

L'algoritmo "J48" presente su "WEKA" corrisponde all'algoritmo "C4.5". L'output può essere rappresentato come un albero decisionale, cioè una struttura definita mediante regole decisionali generate. Tale algoritmo utilizza come criterio di suddivisione di un nodo il "GainRatio", per cui ad ogni split considera l'attributo che garantisce il massimo guadagno di informazione, ovvero la massima riduzione di entropia. Lo split può essere binario o multiplo.

La soglia minima di unità all'interno di ogni nodo (al di sotto della quale non è possibile procedere con la suddivisione) impostata è pari a 100, con lo scopo di ottenere una regola di classificazione non troppo complessa e di facile comprensione.

La regola ricavata ha la seguente rappresentazione grafica.

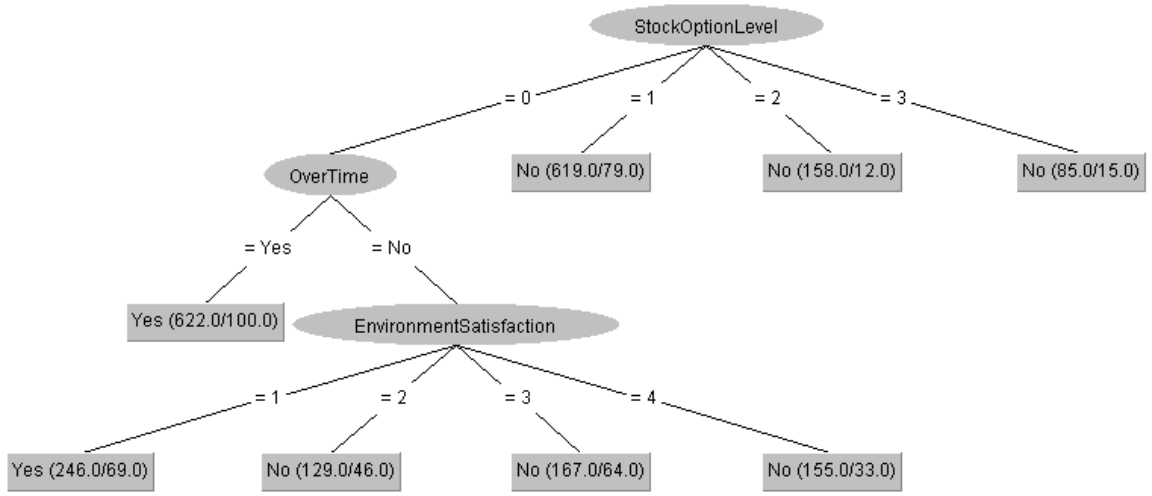


Figura 3.3 Albero derivante dall'algoritmo "J48"

L'albero generato ha 8 nodi terminali e mostra che sono state utilizzate solo le seguenti tre variabili: "StockOptionLevel", "OverTime" e "EnvironmentSatisfaction". Gli altri attributi sono risultati non informativi. La prima variabile utilizzata per splittare il nodo radice è "StockOptionLevel", come era prevedibile in quanto precedentemente nel valutare l'importanza degli attributi era emerso che tale variabile garantiva il massimo valore di "GainRatio" (criterio utilizzato proprio dall'algoritmo "J48"). Dall'albero si evince che coloro che rassegnano le proprie dimissioni sono i dipendenti che:

- presentano un valore di "StockOptionLevel" pari a 0 ed un valore di "OverTime" pari a "Yes";
- presentano un valore di "StockOptionLevel" pari a 0, un valore di "OverTime" pari a "No" ed un livello di "EnvironmentSatisfaction" pari a 1.

Di seguito si riporta l'output ottenuto per l'algoritmo considerato.

```
Time taken to build model: 0.06 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1763               80.8345 %
Incorrectly Classified Instances       418               19.1655 %
Kappa statistic                          0.6062
Mean absolute error                      0.2952
Root mean squared error                  0.3848
Relative absolute error                 60.0636 %
Root relative squared error             77.63   %
Total Number of Instances             2181

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,737    0,137    0,805      0,737    0,770      0,608    0,839     0,767     Yes
                 0,863    0,263    0,810      0,863    0,836      0,608    0,839     0,842     No
Weighted Avg.    0,808    0,208    0,808      0,808    0,807      0,608    0,839     0,809

=== Confusion Matrix ===

   a    b   <-- classified as
 699  249 |    a = Yes
 169 1064 |    b = No
```

Figura 3.4 Output dell'algoritmo "J48"

Il classificatore costruito con l'algoritmo "J48" presenta un valore di accuratezza pari all'80,83%. Le istanze del dataset mal classificate dalla regola costruita sono 418 (si sbaglia a classificare circa il 20% delle unità). Il modello non produrrà spesso buoni risultati poiché non molto robusto: il valore di "Kappa statistic" arriva ad avere un valore pari a 0,606.

3.4 "RandomForest"

Il "RandomForest" è un algoritmo di classificazione rappresentato da più alberi decisionali casuali addestrati sullo stesso insieme di dati.

Seguono i risultati dell'algoritmo applicato.

```
Time taken to build model: 1.23 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1951               89.4544 %
Incorrectly Classified Instances       230               10.5456 %
Kappa statistic                          0.7836
Mean absolute error                      0.1875
Root mean squared error                  0.2852
Relative absolute error                 38.1503 %
Root relative squared error             57.5239 %
Total Number of Instances             2181

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,842    0,065    0,909      0,842    0,874      0,785    0,950     0,952     Yes
               0,935    0,158    0,885      0,935    0,909      0,785    0,950     0,943     No
Weighted Avg.  0,895    0,118    0,895      0,895    0,894      0,785    0,950     0,947

=== Confusion Matrix ===

   a    b   <-- classified as
 798  150 |    a = Yes
  80 1153 |    b = No
```

Figura 3.5 Output dell'algoritmo "RandomForest"

Il modello ottenuto consente di avere un'alta accuratezza poiché questa raggiungere un valore pari a 0,895. Le istanze mal classificate sono 230, di cui 150 sono dipendenti che hanno dato le proprie dimissioni e 80 sono dipendenti che non hanno abbandonato l'azienda. Questo classificatore è sufficientemente robusto con valore di "Kappa statistic" quasi pari a 0,8. Per quanto riguarda i valori di "F-Measure" (misura che riassume "Precision" e "Recall"), si osserva che questi sono abbastanza elevati per entrambe le etichette di classe.

3.5 Scelta del classificatore

Si prosegue adesso a confrontare, attraverso la tabella che segue, i risultati ottenuti, al fine di selezionare il classificatore che potrà essere utilizzato in futuro per classificare nuove istanze, così da poter prevenire e controllare nel miglior modo possibile il fenomeno delle dimissioni da parte dei dipendenti.

Classificatore	Kappa statistic	Accuracy
NaiveBayes	0,5907	0,7969
AdaBoostM1	0,6369	0,8207
J48	0,6062	0,8083
RandomForest	0,7836	0,8945

Figura 3.6 Tabella riassuntiva dei valori di "Kappa statistic" e "Accuracy" per gli algoritmi considerati

I primi tre classificatori sono molti simili tra loro in termini di valori delle statistiche esaminate. Il classificatore da preferire è, senza dubbio, quello generato tramite l'algoritmo "RandomForest", principalmente perché il più robusto, ovvero con valore di "Kappa statistic" più alto (0,7836), ed in secondo luogo perché è quello che presenta il valore di "Accuracy" più elevato (89,45%), permettendo di conseguenza di avere il più basso numero di casi mal classificati, rispetto agli altri modelli matematici creati. Infine, per quanto riguarda il parametro di valutazione relativo alla velocità del classificatore, il "RandomForest" ha richiesto un tempo complessivo pari a 1.23 secondi.

Al fine di ottenere ulteriori conferme a questa scelta, è possibile utilizzare, come strumento di valutazione, il "Clustering", task di apprendimento non supervisionato. Infatti, apprendimento supervisionato (classificazione) e non supervisionato possono essere considerati complementari, in quanto ognuno dei due approcci consente di valutare la tecnica opposta.

4. Clustering

Al fine di valutare l'appartenenza di un soggetto ad uno specifico cluster avente determinate proprietà, si impiegheranno algoritmi di apprendimento non supervisionato, in particolare: "K-means" e "Fartherst First".

4.1 "SimpleKMeans"

L'algoritmo "K-means" fa parte degli algoritmi di clustering partizionali e il primo passo per la sua applicazione è la determinazione a priori del numero di cluster in cui si vogliono ripartire i soggetti, in questo caso 2.

Inoltre, si seleziona l'opzione "Classes to cluster evaluation" presente su "WEKA", che permette di avere un raffronto diretto con i risultati già ottenuti mediante l'impiego degli algoritmi di classificazione considerati precedentemente.

Il risultato dell'applicazione di tale algoritmo ha prodotto il seguente output.

```
Number of iterations: 6
Within cluster sum of squared errors: 6436.930267580921

Initial starting points (random):

Cluster 0: 0.857143,Travel_Rarely,4,3,4,0.859347,No,0,0.725,2,0.5,0.411765
Cluster 1: 0.285714,Travel_Rarely,1,3,3,0.059136,Yes,0,0.175,3,0.125,0.058824

Missing values globally replaced with mean/mode

Final cluster centroids:
                                        Cluster#
Attribute                 Full Data             0             1
                           (2181.0)      (1085.0)      (1096.0)
=============================================================
Age                          0.4231        0.4673        0.3792
BusinessTravel        Travel_Rarely Travel_Rarely Travel_Rarely
EnvironmentSatisfaction         1             4             1
JobInvolvement                  3             3             3
JobSatisfaction                 3             4             3
MonthlyIncome                0.2598        0.3125        0.2077
OverTime                        No            No           Yes
StockOptionLevel                 0             0             0
TotalWorkingYears            0.2563        0.3014        0.2118
WorkLifeBalance                 3             3             3
YearsAtCompany               0.1595        0.1964        0.1229
YearsWithCurrManager         0.2189        0.2727        0.1656

Time taken to build model (full training data) : 0.34 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      1085 ( 50%)
1      1096 ( 50%)


Class attribute: Attrition
Classes to Clusters:

   0   1  <-- assigned to cluster
 216 732 | Yes
 869 364 | No

Cluster 0 <-- No
Cluster 1 <-- Yes

Incorrectly clustered instances :       580.0    26.5933 %
```

Figura 4.1 Applicazione dell'algoritmo "SimpleKMeans" considerando due cluster

Dall'output sembrerebbe che le variabili "BusinessTravel", "JobInvolvement", "StockOptionLevel" e "WorkLifeBalance" non siano significative al fine di raggruppare i soggetti in due cluster. Tuttavia, dagli algoritmi di classificazione, è stato possibile accertare che la variabile "StockOptionLevel" svolge un importante ruolo nella classificazione dei soggetti, in particolare per quanto riguarda il classificatore ottenuto dal "J48".

L'algoritmo ha effettuato 6 iterazioni con errore quadratico pari a 6436.93. Inoltre, il 50% delle unità ricade rispettivamente in ciascun cluster, con un tasso di errata classificazione pari al 26.56%, ovvero 580 unità non sono state correttamente classificate.

4.2 "FarthestFirst"

Si procede ora alla valutazione delle regole di classificazione ottenute mediante l'impiego dell'algoritmo "FarthestFirst". Si tratta anch'esso di un algoritmo di clustering partizionale all'interno del quale però, a differenza del "K-means", il centroide di ogni cluster è selezionato come il punto più lontano dai centri dei cluster esistenti.

Di seguito si riporta l'output ottenuto applicando l'algoritmo in questione.

```
=== Clustering model (full training set) ===


FarthestFirst
==============

Cluster centroids:

Cluster 0
        0.40476190476190477 Travel_Rarely 4 3 3 0.07345971563981042 No 1 0.025 3 0.025 0.0
Cluster 1
        0.8095238095238095 Non-Travel 1 2 3 0.9509741969457609 Yes 0 0.825 4 0.825 0.7058823529411765



Time taken to build model (full training data) : 0.05 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      2000 ( 92%)
1       181 (  8%)

Class attribute: Attrition
Classes to Clusters:

    0    1  <-- assigned to cluster
  841  107 | Yes
 1159   74 | No

Cluster 0 <-- No
Cluster 1 <-- Yes

Incorrectly clustered instances :       915.0    41.9532 %
```

Figura 4.2 Applicazione dell'algoritmo "FarthestFirst" considerando due cluster

Secondo tale algoritmo di clustering, solo l'attributo "JobSatisfaction" non risulta significativo nell'identificazione dei due cluster, in quanto assume il medesimo valore in

entrambi i cluster, mentre le altre variabili risultano significative. Per quanto concerne la classificazione effettuata dopo l'assegnazione delle classi ai due cluster (composti da 2000 e 181 unità), si può notare che il numero di istanze non classificate correttamente è ora pari a 915, con un tasso di errata classificazione del 41.95%, maggiore rispetto a quello riscontrato applicando l'algoritmo "K-means".

5. Conclusioni

La soluzione proposta, raggiunta dopo l'impiego di diversi algoritmi di mining, rappresenta un valido strumento che consente di perseguire lo scopo delle analisi effettuate, ovvero quello di ridurre quanto più possibile il numero di lavoratori dipendenti che rassegnano le proprie dimissioni e, quindi, di sviluppare e rafforzare l'Employee Retention dell'azienda.

È altresì importante osservare che i risultati ottenuti sono sicuramente migliorabili attraverso analisi più approfondite e grazie ad una quantità di dati superiore a quella a disposizione.