Telco Customer Churn

Programma orientato alla fidelizzazione dei clienti

Il dataset in esame, contenente dati relativi ai clienti di un'azienda di telecomunicazioni, consente un'analisi esplorativa delle informazioni personali e contrattuali dei clienti, allo scopo di individuare programmi di incentivazione e promozione focalizzati sulla fidelizzazione della clientela. Il progetto è vòlto alla costruzione di un meccanismo predittivo che possa individuare quali clienti potrebbero essere più soggetti a 'churn', cioè ad abbandonare l'azienda e, dunque, ad indicare quali segmenti della clientela potrebbe essere fruttuoso sottoporre a misure preventive di promozione.

Nel dataset, ad ogni riga corrisponde un cliente, per un totale di 7043 righe, mentre nelle colonne vengono specificati gli attributi di ogni cliente, per un totale di 21 colonne. Di queste, cinque sono riconducibili alle informazioni personali del cliente: 'customerId', che corrisponde all'identificativo univoco del cliente, 'gender', per il sesso, 'SeniorCitizen', 'Partner', 'Dependents' che indicano se il cliente è, rispettivamente, un cliente 'senior', se ha un partner o se ha dei dipendenti. Le restanti colonne si riferiscono agli accordi contrattuali, dunque ai servizi utilizzati ('PhoneService', 'MultipleLines', 'InternetService', 'OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport', 'StreamingTV', 'StreamingMovies'), alla tipologia di contratto ('Contract') e ai pagamenti ('PaperlessBilling', 'PaymentMethod', 'MonthlyCharges', 'TotalCharges'). Le colonne 'Churn' e 'tenure' considerano rispettivamente se il cliente abbia abbandonato la compagnia e i mesi effettivi di permanenza.

Nel seguente report, la prima sezione verrà dedicata al Data Understanding e al pre-processing (Data Manipulation), in cui viene fornita una descrizione più approfondita delle variabili in esame e delle loro relazioni interne. Verrà valutata la qualità e completezza dei dati, la presenza di valori mancanti e di outliers che verranno gestiti di conseguenza. Verrà, inoltre, valutato il valore informativo delle feature, per poter effettuare una selezione di quelle più rilevanti.

La seconda sezione sarà dedicata all'analisi descrittiva del dataset attraverso metodi di clustering. Verranno, infatti, applicati diversi algoritmi di clustering, di cui verrà effettuato uno studio valutativo per accertarne/confermarne la validità.

La terza parte sarà invece dedicata ai metodi predittivi attraverso la classificazione. Tenendo come target il 'churn', verrà, dunque, allenato un modello per la predizione dell'abbandono, su cui verrà effettuata un'operazione di tuning mediante cross validation e che verrà in seguito valutato con un test set. Si cercherà, inoltre, di valutare le performance di altri classificatori, in particolare di metodi di ensemble.

La quarta sezione, per completare l'analisi descrittiva del dataset, si concentrerà, invece, sul pattern mining. La ricerca di pattern d'interesse all'interno del dataset verrà affiancata dall'estrazione di regole associative e da considerazioni di tipo qualitativo.

1. Data Understanding and Manipulation

Come analisi preliminare, abbiamo visto di che tipo erano le istanze di ogni feature (*Fig.1*), scoprendo che gli attributi del dataset in esame erano per la maggior parte categorici. Infatti, gli unici attributi numerici sono risultati essere la 'tenure', il 'MonthlyCharges' e il 'SeniorCitizen'. La natura numerica di quest'ultimo attributo è però inconsistente con le informazioni che essa porta, che sono informazioni da iscrivere all'insieme dei categorici. Un'ulteriore inconsistenza è stata individuata nel tipo dell'attributo 'TotalCharges', che risulta essere di tipo stringa, nonostante si tratti di una variabile numerica. Inoltre, per undici clienti il valore risulta essere una stringa vuota; pertanto, in fase di conversione in numerica, queste sono diventate valori nulli.

variable	type	domain	any_null
customerID	<class 'str'>	[0002-ORFBO, 0003-MKNFE, 0004-TLHLJ, 0011-IGKF...	False
gender	<class 'str'>	[Female, Male]	False
SeniorCitizen	<class 'numpy.int64'>	[0, 1]	False
Partner	<class 'str'>	[No, Yes]	False
Dependents	<class 'str'>	[No, Yes]	False
tenure	<class 'numpy.int64'>	[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13,...	False
PhoneService	<class 'str'>	[No, Yes]	False
MultipleLines	<class 'str'>	[No, No phone service, Yes]	False
InternetService	<class 'str'>	[DSL, Fiber optic, No]	False
OnlineSecurity	<class 'str'>	[No, No internet service, Yes]	False
OnlineBackup	<class 'str'>	[No, No internet service, Yes]	False
DeviceProtection	<class 'str'>	[No, No internet service, Yes]	False
TechSupport	<class 'str'>	[No, No internet service, Yes]	False
StreamingTV	<class 'str'>	[No, No internet service, Yes]	False
StreamingMovies	<class 'str'>	[No, No internet service, Yes]	False
Contract	<class 'str'>	[Month-to-month, One year, Two year]	False
PaperlessBilling	<class 'str'>	[No, Yes]	False
PaymentMethod	<class 'str'>	[Bank transfer (automatic), Credit card (autom...	False
MonthlyCharges	<class 'numpy.float64'>	[18.25, 18.4, 18.55, 18.7, 18.75, 18.8, 18.85,...	False
TotalCharges	<class 'numpy.float64'>	[18.9, 19.0, 19.05, 19.15, 19.25, 19.45, 19.55...	True
Churn	<class 'str'>	[No, Yes]	False

Fig. 1 Variabili in esame e loro tipo.

Dalla panoramica mostrata sopra, è possibile vedere come il valore booleano *True* nell'ultima colonna, segnali la presenza di valori nulli nella feature 'TotalCharges'. Al fine di risolvere questa e le altre inconsistenze riscontrate nel dataset con questa indagine preliminare, abbiamo per prima cosa rimappato la feature 'SeniorCitizen' da binaria numerica a categorica, sostituendo 0 e 1 con 'No' e 'Yes'.

Come è possibile osservare (*Fig. 2*), l'unica feature con valori nulli è 'TotalCharges'. Nonostante l'esiguità del numero, abbiamo controllato se ci fosse un pattern per l'occorrenza del valore *NaN* in questa feature e abbiamo visto che essa era presente ogni qual volta i valori della feature 'tenure' assumevano valore 0. Per questo motivo, in un primo momento abbiamo

trasformato i valori *NaN* presenti nel nostro dataset in valori 0. In seguito, però, per semplificare le analisi, abbiamo eliminato la colonna 'TotalCharges', in quanto contenente un'informazione ridondante, ovvero i valori risultano essere approssimativamente il prodotto dei valori delle colonne 'tenure' e 'MonthlyCharges'.

```
Out[8]: gender            0
        SeniorCitizen     0
        Partner           0
        Dependents        0
        tenure            0
        PhoneService      0
        MultipleLines     0
        InternetService   0
        OnlineSecurity    0
        OnlineBackup      0
        DeviceProtection  0
        TechSupport       0
        StreamingTV       0
        StreamingMovies   0
        Contract          0
        PaperlessBilling  0
        PaymentMethod     0
        MonthlyCharges    0
        TotalCharges     11
        Churn             0
        dtype: int64
```

Fig. 2 Valori nulli.

Out[11]:

id	TotalCharges	tenure
488	NaN	0
753	NaN	0
936	NaN	0
1082	NaN	0
1340	NaN	0
3331	NaN	0
3826	NaN	0
4380	NaN	0
5218	NaN	0
6670	NaN	0
6754	NaN	0

Fig. 3 Confronto fra i valori nulli di TotalCharges e tenure.

Proseguendo la nostra indagine preliminare, abbiamo ottenuto una proiezione delle distribuzioni di tutte le nostre feature. È facilmente osservabile come, all'interno della nostra feature target, 'Churn', il numero dei 'Churn=yes' è di poco superiore ad un quarto del nostro dataset (*Fig. 4*).

Fig. 4 Distribuzione del Churn.

È stato, inoltre, interessante notare come all'interno della feature 'tenure' ci sia un'alta concentrazione di valori nell'intervallo della prima metà del bin 0-20 e nella seconda metà dell'ultimo bin, indicazione che i clienti del dataset in esame tendono o a rescindere il contratto dopo i primi mesi o a fidelizzarsi (*Fig.5)*.

Altra distribuzione interessante ci è parsa quella dei valori all'interno della feature 'MonthlyCharges', che come è possibile notare, pare essere molto sbilanciata sui valori più bassi, segno che una buona parte dei clienti presenti nel dataset gode di tariffe mensili molto basse.

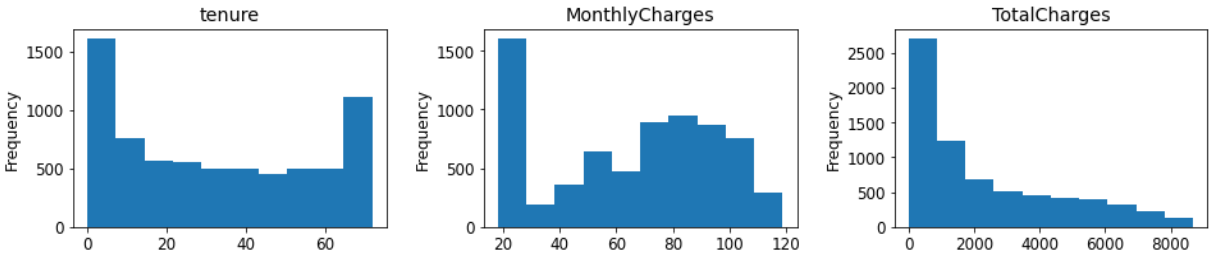

Fig. 5 Distribuzione delle variabili numeriche.

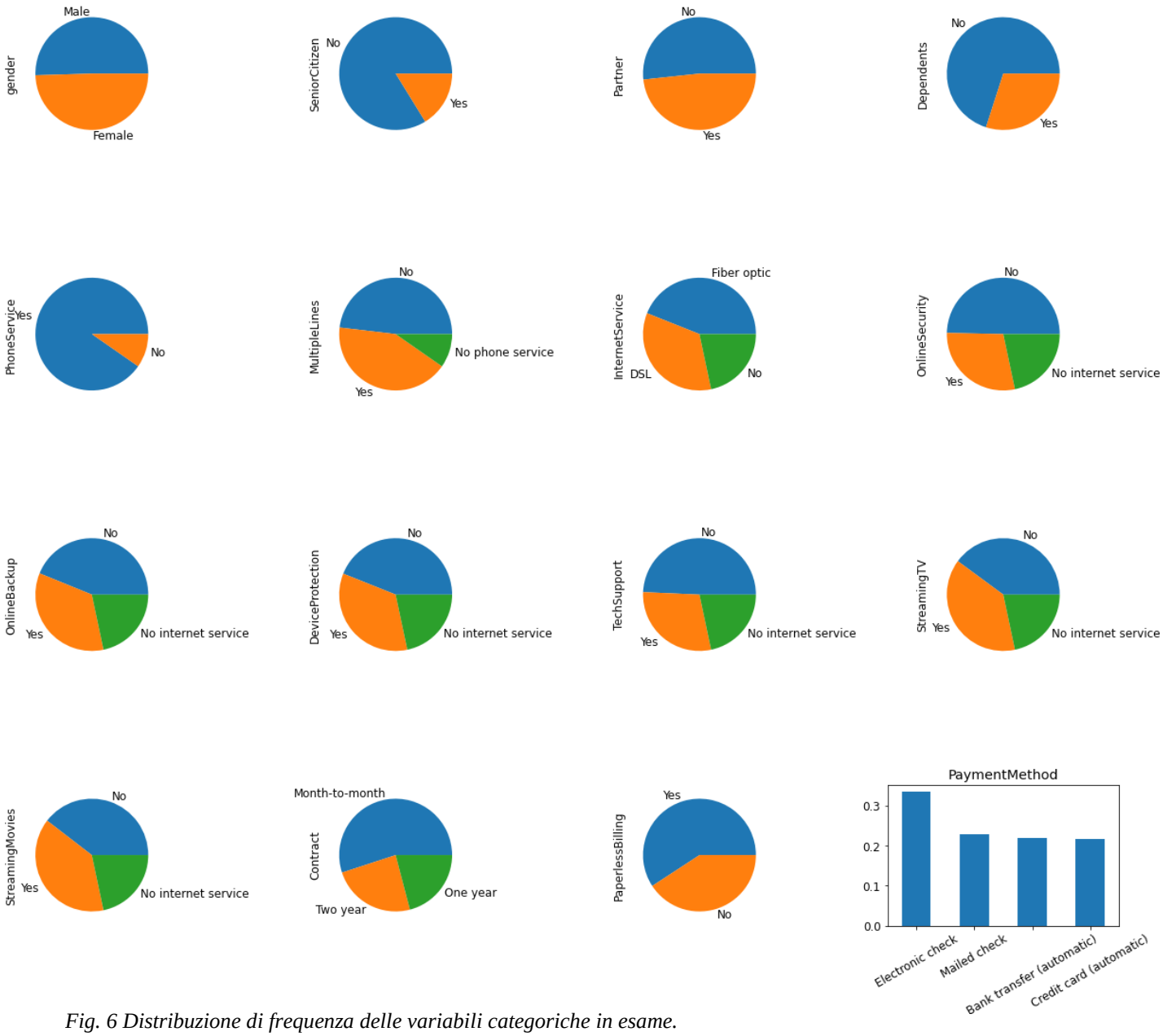

Fig. 6 Distribuzione di frequenza delle variabili categoriche in esame.

Abbiamo in seguito circoscritto l'analisi esplorativa alla nostra feature target, mettendola in relazione con tutte le altre feature presenti nel nostro dataset. Per quanto riguarda le feature numeriche, abbiamo visualizzato la variazione di densità di 'MonthlyCharges' e di 'tenure' per 'Churn' e, successivamente, proiettato lo *scatter plot* del 'Churn' con la 'tenure' sull'asse delle ascisse e 'MonthlyCharges' sull'asse delle ordinate . Nel caso di 'MonthlyCharges' e 'tenure' (*Fig.7*), l'intuizione che sovviene è quella che gli abbandoni siano più numerosi laddove il piano tariffario mensile della compagnia telefonica è più elevato. La *Fig. 8*, invece, mostra un picco di densità davvero significativo per quanto riguarda il valore 'Churn=yes' nell'intervallo 0-20 della tenure, che si ricollega a quello che è possibile osservare in *Fig. 5*. Questo dato ci fa inferire che i clienti abbandonano generalmente nei primi mesi dell'attivazione del contratto, mentre la curva si abbassa precipitosamente con l'aumentare dei mesi.

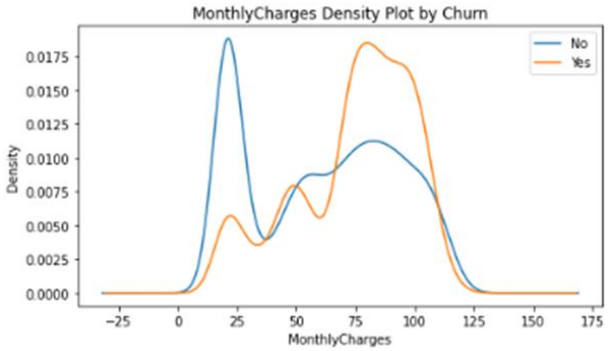

Fig. 7 Density plot di MonthlyCharges per Churn.

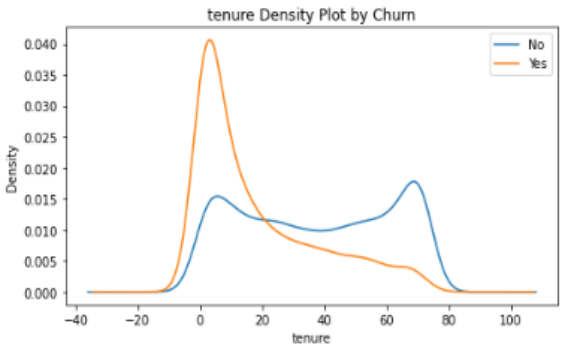

Fig. 8 Density plot di tenure per Churn.

La figura 10 conferma le intuizioni precedenti, con i punti rossi rappresentanti i 'Churn=yes' che formano un agglomerato importante nei primi intervalli della feature 'tenure' e con l'aumentare della feature 'MonthlyCharges'.

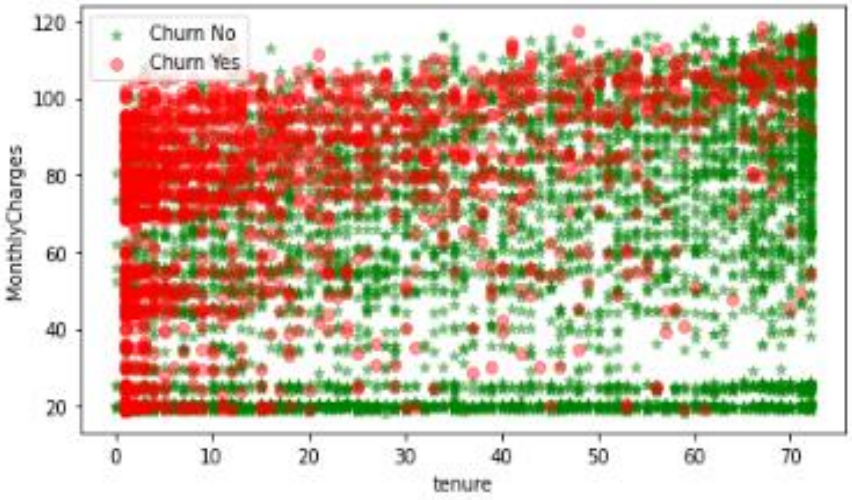

Fig. 9 Scatter plot di tenure e MonthlyCharges per Churn.

Infine, abbiamo visualizzato la relazione tra la nostra feature target e il resto delle feature presenti all'interno del dataset, come è possibile vedere in *Fig.10*. Le feature più significative che sono emerse sono quelle tra l'abbandono del cliente ('Churn=yes'*)* e i già citati servizi offerti dalla compagnia È facilmente osservabile che, ad esempio nel caso del servizio 'OnlineSecurity', la concentrazione dei 'Churn=yes' è molto alta in coloro che non hanno usufruito di tale servizio. E questo pattern si ripete per quasi tutti i servizi presenti. Una situazione differente invece si presenta per il servizio 'InternetService', che vede una concentrazione molto alta di 'Churn=yes' in corrispondenza dell'opzione 'fiber optic'. Per concludere, anche la cross-tabulation tra il tipo di contratto offerto dalla compagnia telefonica e l'abbandono degli utenti risulta essere molto informativa. Possiamo vedere, infatti, come la maggior parte degli 'Churn=yes' abbiano optato per l'opzione contrattuale 'Month-to-month'.

Fig. 10 Feature del dataset in relazione al Churn.

2. Clustering

2.1 Clustering per variabili numeriche

Data la natura delle feature del dataset, contenente solo tre variabili numeriche, abbiamo effettuato un primo tentativo di clustering su quest'ultime, tenendo in considerazione, dunque, 'Monthly Charges' e 'tenure'.

id	tenure	MonthlyCharges
0	1	29.85
1	34	56.95
2	2	53.85
3	45	42.30
4	2	70.70

Come è possibile osservare dal dataframe, i valori delle feature operano su scale diverse e, prima di procedere con gli algoritmi di clustering, è stato necessario riportare i valori su una scala equiparabile. Per fare questo, abbiamo applicato il **MinMaxScaler**.

Fig. 11 Variabili numeriche.

Il primo algoritmo di cluster utilizzato è stato il **K-Means**. Per trovare un numero ragionevole di centroidi da usare come parametro k, abbiamo una costruito una curva del **SSE** e una curva della **Silhouette** che indicassero il ridursi dell'errore all'aumentare dei centroidi.

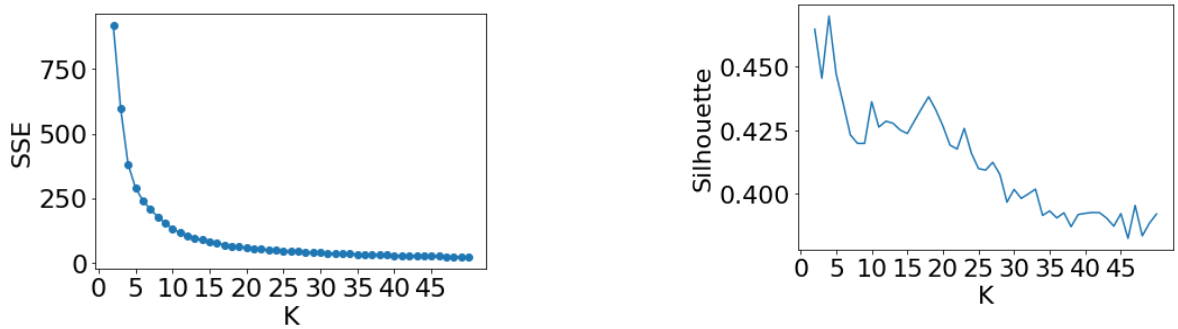

Fig. 12 Curva SSE-cluster e curva Silhouette-cluster.

Visti i risultati delle curve, abbiamo stabilito il numero di centroidi a 4. La distribuzione dei cluster risultante è leggermente sbilanciata verso uno dei cluster, che raccoglie il 32,19 % delle osservazioni, mentre gli altri tre ne hanno rispettivamente il 27,26%, il 24,08 % e il 16,47% nel cluster meno numeroso.

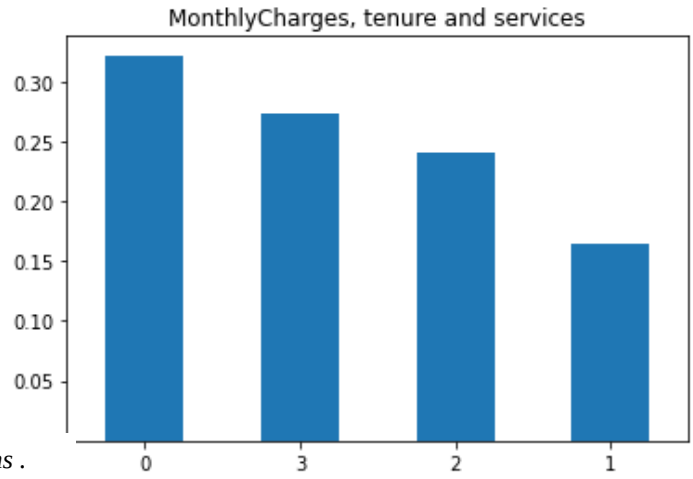

Fig. 13 Distribuzione dei cluster con K-means .

Il modello sviluppato con K-Means restituisce un score di silhouette dello 0,47, evidenziando una discreta coesione nei cluster.

Per facilitare la visualizzazione, abbiamo plottato le due feature numeriche più significative ('Monthly Charges' e 'tenure') in uno scatter plot, in cui abbiamo colorato i cluster con colori diversi ed evidenziato la posizione dei quattro centroidi.

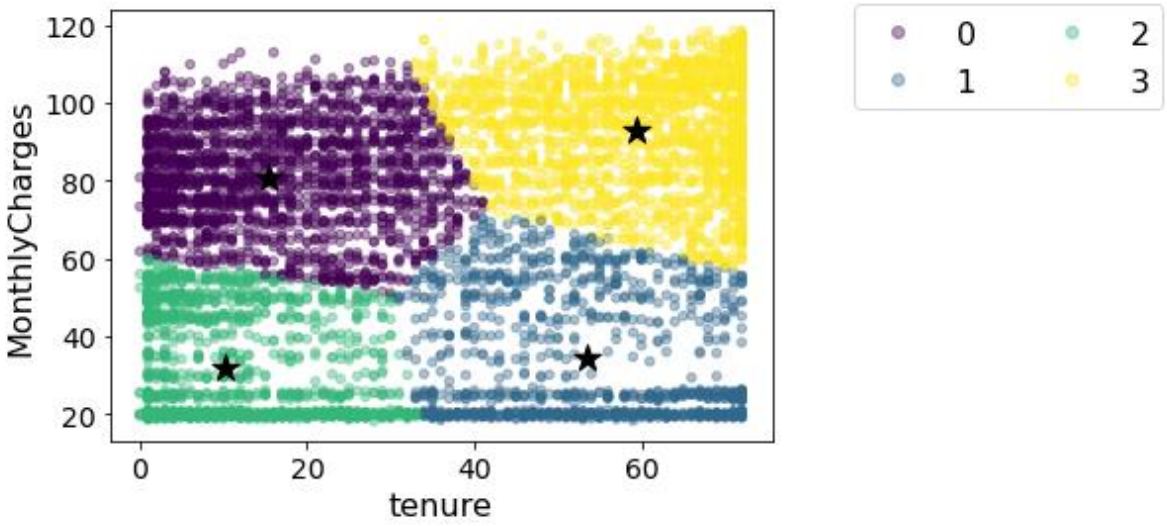

Fig. 14 Scatter plot delle variabili numeriche colorate per cluster.

Un'altra visualizzazione che abbiamo utilizzato riporta i quattro cluster individuati al 'Churn', la variabile target del dataset. Dividendo 'Churn Yes' da 'Churn No', è possibile osservare che una parte consistente dei clienti che hanno abbandonato l'azienda, fanno parte del cluster 0 che, come risulta visibile nello scatter plot precedente, rientra in quella sezione di clienti con una spesa mensile medio-alta e una permanenza medio-bassa.

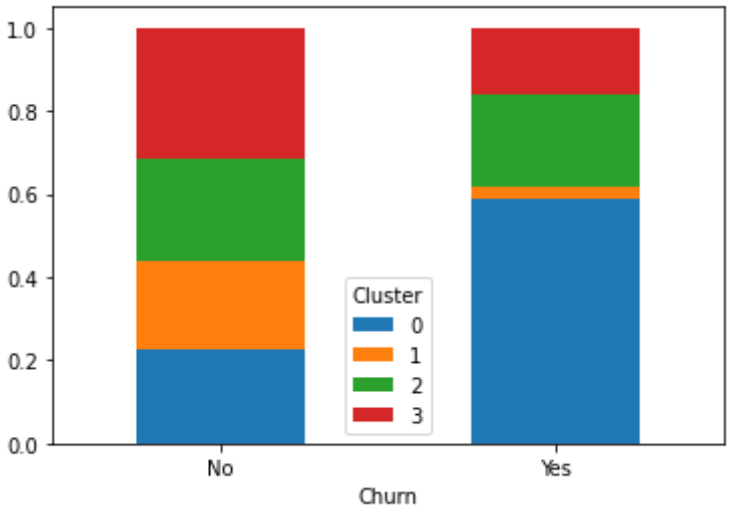

Fig. 15 Divisione in cluster in relazione al Churn.

Oltre al K-Means, abbiamo provato ad applicare altri algoritmi di clustering sulle variabili numeriche, che si sono, però, rivelati meno efficaci. Come atteso, considerata la natura dell'algoritmo e la disposizione dei punti del nostro dataset, il **DBSCAN** è stato l'algoritmo più inefficiente, formando numerosi cluster di piccolissime dimensioni e uno più grande contenente il 97% circa delle osservazioni. Il **gerarchico** si è rivelato ugualmente poco efficiente, con un coefficiente di silhouette del 0,36.

2.2 Clustering per variabili categoriche

Per gestire le variabili categoriche, che costituiscono la maggior parte delle feature del dataset, abbiamo invece utilizzato il **K-Modes**, un algoritmo di clustering che sfrutta il numero di match fra categorie nei dati. Dopo aver selezionato le feature più significative per il 'Churn' tramite i plot dei cross-tab in *Fig. 10*, abbiamo generato i grafici del costo (equivalente dell'SSE) e della silhouette al variare della variabile K. E proprio sulla base di questi grafici abbiamo deciso che un buon compromesso per la variabile K in termini di complessità e 'bontà' del clustering è il valore 5.

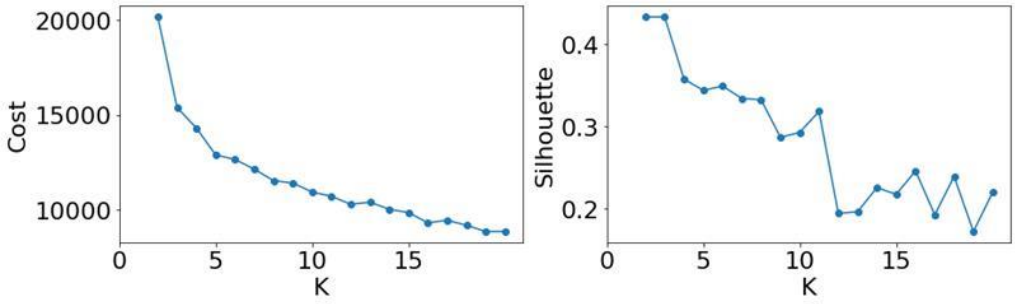

Fig. 16 Costo e silhouette su numero di cluster.

Per il calcolo della silhouette è stato necessario generare la matrice delle distanze. Questa è stata calcolata utilizzando lo stesso criterio con cui l'algoritmo KModes ha riportato prestazioni migliori: l'indice di Simple Matching.

Applicando 5 come numero di cluster all'algoritmo abbiamo ottenuto la seguente distribuzione:

Fig. 17 Distribuzione dei cluster con K-Modes.

Com'è possibile osservare dal bar chart, i cluster non risultano essere perfettamente omogenei. Per aumentare la comprensione dei risultati, dunque, abbiamo messo in relazione i cluster con le due variabili numeriche mediante scatter plot, ed emerge una distribuzione dei cluster in termini di 'MonthlyCharges' e 'tenure' abbastanza distinguibile e separata.

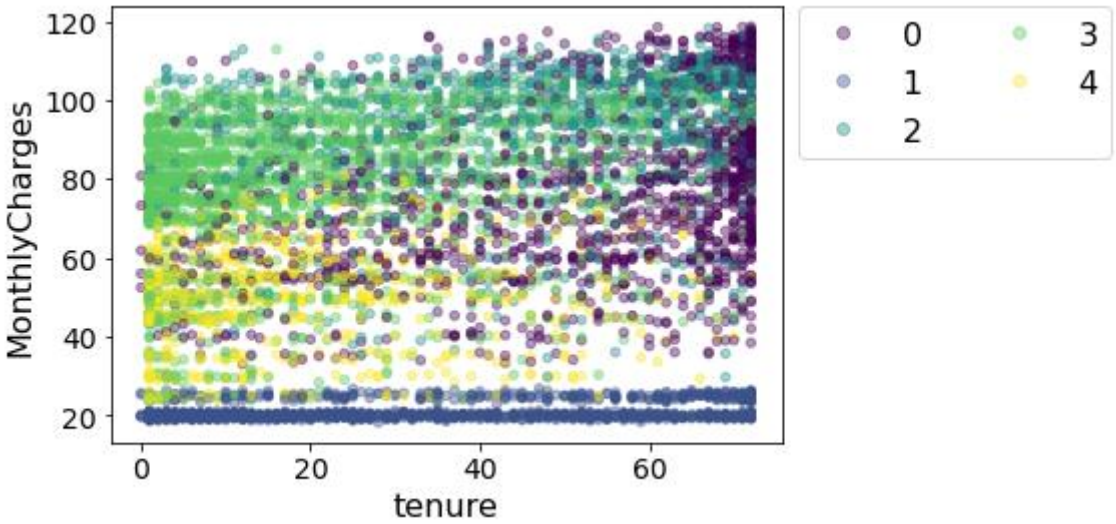

Fig. 18 Scatter plot delle variabili numeriche con evidenziati i cluster ottenuti con Kmodes.

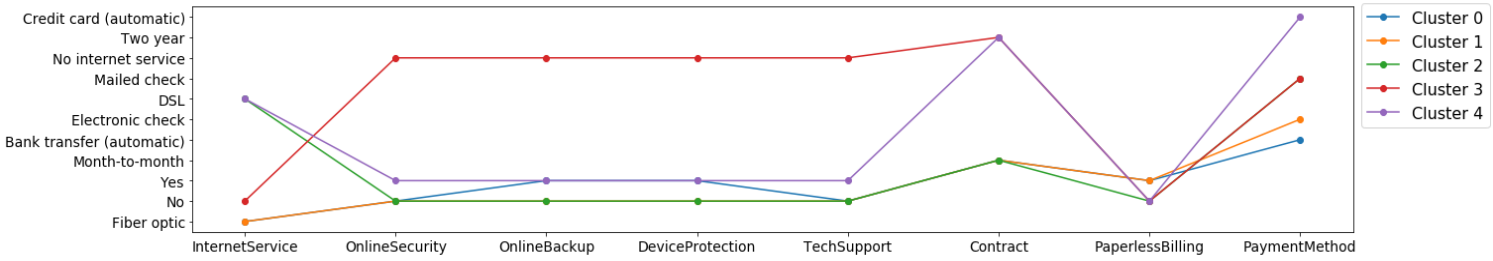

Fig. 19 Parallel Coordinates dei centroidi.

È stato particolarmente informativo plottare il parallel coordinates graph (il grafico dei centroidi), in quanto fornisce un'immagine caratterizzante dei cluster attraverso l'identificazione dei centroidi (*Fig.19*).

Come ultima visualizzazione, abbiamo evidenziato la partizione in cluster, suddivisa per 'Churn Yes' e 'Churn No'(*Fig.20*). Come è visibile dal bar chart, si evidenzia che una maggior parte dei clienti che hanno abbandonato l'azienda corrispondono al cluster 1, che risulta anche essere il cluster più numeroso.

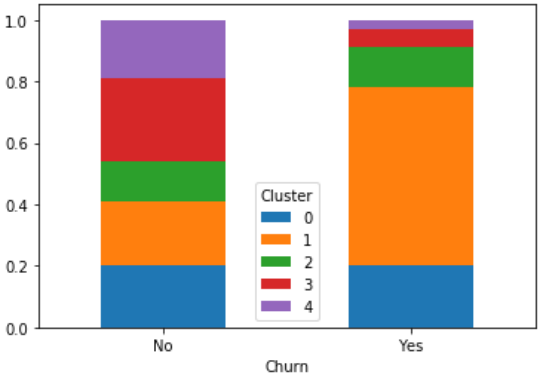

Fig. 20 Partizione in cluster per Churn.

3. Classification

La classificazione si è rivelata un'ottima soluzione per il problema in analisi: individuare i clienti che abbandonano la compagnia e capire quali sono le variabili chiavi nel determinare se un cliente rimarrà o sceglierà di passare ad un competitor. Abbiamo selezionato come valore da predire il 'Churn', diventato la nostra label target. I modelli di classificazione che abbiamo deciso di provare sono stati due: DecisionTree e RandomForest.

Le feature sono state preparate per la classificazione trasformando in *dummies* le variabili categoriche non binarie, mentre le features binarie sono state trasformate in 0 ed 1 attraverso la classe *LabelEncoder* di Sklearn e senza modificare le due variabili numeriche. Il dataset è stato quindi diviso in training e test set, attraverso l'uso della funzione *train_test_split* della libreria Sklearn, che assegna in modo randomico le osservazioni del dataset nei due gruppi con le proporzioni del 70% per il training set e del 30% per il test set. Abbiamo, inoltre, specificato il parametro 'stratify' per mantenere la distribuzione della variabile target sia nel trainig che nel test.

3.1 Decision Tree

Il metodo **DecisionTree** prevede la scelta di diversi iperparametri e, per poter valutare quali potessero essere i più idonei, abbiamo utilizzato una funzione in grado di ricercare

combinazioni ottimali di questi parametri e ottenere così il miglior stimatore. Data in input una lista di possibili valori di iperparametri, la classe *RandomizedSearchCV* di Sklearn, è riuscita a individuare il miglior stimatore mediante la generazione di diversi classificatori e la loro validazione con cross-validation.

```
param_list = {
    'max_depth': [None] + list(np.arange(2, 20)),
    'min_samples_split': [2, 5, 10, 20, 30, 50, 100],
    'min_samples_leaf': [1, 5, 10, 20, 30, 50, 100],
    'criterion': ['gini', 'entropy']}
```

Fig.21 Lista di possibili iperparametri.

Abbiamo quindi fittato il modello sul training set e osservato quali feature avevano un'importanza superiore allo 0 (*Fig.22*).

Grazie all'utilizzo di *feature_selection.SelectFromModel* di Sklearn siamo stati in grado di selezionare solo le feature al di sopra di una soglia di importanza pari allo 0.01, così da poter rendere un po' più generalizzato il nostro modello. A questo punto, tenendo in considerazione solo le feature selezionate, abbiamo diviso nuovamente il dataset iniziale in training e test set e abbiamo provato a riallenare un modello con il Decision Tree Classifier.

features	importance
Contract=Month-to-month	0.507003
tenure	0.186186
InternetService=Fiber optic	0.178297
MonthlyCharges	0.028517
PaymentMethod=Electronic check	0.026559
OnlineSecurity=No	0.024440
OnlineBackup=No	0.013876
Contract=Two year	0.010646
StreamingMovies=Yes	0.009623
MultipleLines=No	0.008760
Contract=One year	0.002707
SeniorCitizen	0.002345
gender	0.000888
PaymentMethod=Bank transfer (automatic)	0.000153
OnlineBackup=No internet service	0.000000
MultipleLines=Yes	0.000000
PaperlessBilling	0.000000
Dependents	0.000000
Partner	0.000000
PhoneService	0.000000
PaymentMethod=Mailed check	0.000000
InternetService=DSL	0.000000
PaymentMethod=Credit card (automatic)	0.000000
InternetService=No	0.000000
OnlineSecurity=No internet service	0.000000
OnlineBackup=Yes	0.000000
OnlineSecurity=Yes	0.000000
StreamingMovies=No internet service	0.000000
StreamingMovies=No	0.000000
StreamingTV=Yes	0.000000
MultipleLines=No phone service	0.000000
StreamingTV=No	0.000000
TechSupport=Yes	0.000000
TechSupport=No internet service	0.000000
TechSupport=No	0.000000
DeviceProtection=Yes	0.000000
DeviceProtection=No internet service	0.000000
DeviceProtection=No	0.000000
StreamingTV=No internet service	0.000000

Fig.22 Feature importance.

Abbiamo così generato l'albero di decisione che tiene in considerazione solo le feature selezionate: 'Contract=Month-to-month', 'tenure', 'InternetService=Fiber optic', 'MonthlyCharges', 'PaymentMethod=Electronic check', 'OnlineSecurity=No', 'OnlineBackup=No', 'Contract=Two Year'.

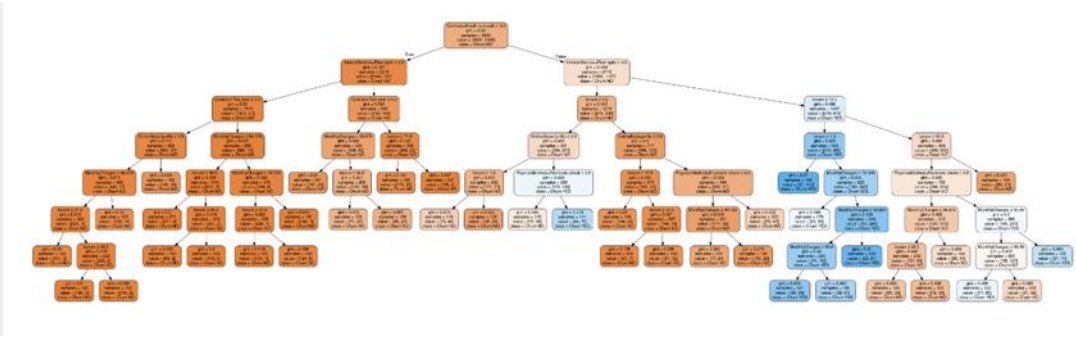

Fig.23 Albero decisionale.

Per la valutazione del DecisionTree Classifier abbiamo analizzato i valori di accuracy e F1 score per training e test set, valutandoli entrambi soddisfacenti intorno allo 0.8.

```
Train Accuracy 0.8016227180527383
Train F1-score [0.86886565 0.5928393 ]

Test Accuracy 0.7974443918599148
Test F1-score [0.86683261 0.5770751 ]
              precision    recall  f1-score   support

           0       0.84      0.90      0.87      1552
           1       0.65      0.52      0.58       561

    accuracy                           0.80      2113
   macro avg       0.74      0.71      0.72      2113
weighted avg       0.79      0.80      0.79      2113
```

Fig.24 Metriche di valutazione.

La ConfusionMatrix (*Fig. 25*) ci ha poi permesso di osservare il modello in termini di divisione numerica tra TP, FP, TN, FN, evidenziando quanto il modello sia però in grado di individuare più facilmente i clienti che non abbandonano la compagnia, i 'Churn=No'. La ROC curve ci ha altresì permesso di confermare un'area sotto la curva di 0.7, dato valutato positivamente in termini di capacità predittiva del modello (*Fig.26*).

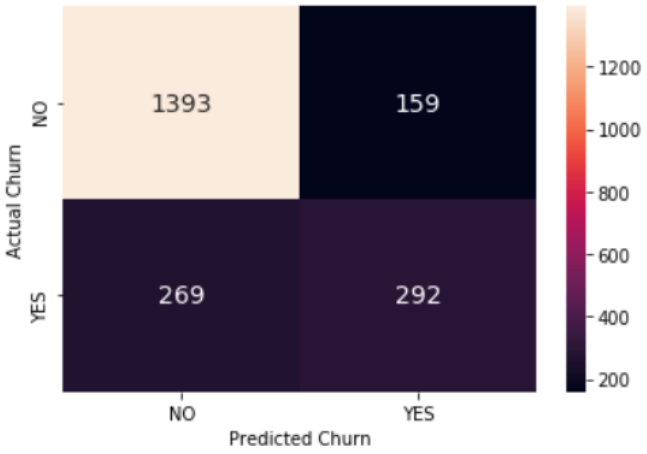

Fig.25 Confusion matrix per il Decision Tree.

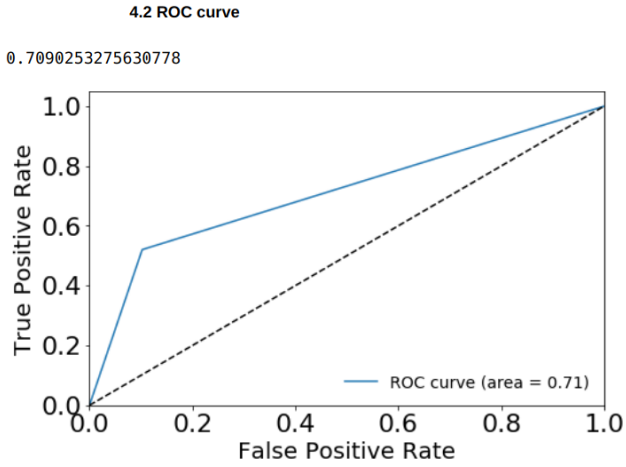

Fig.26 Curva ROC per il Decision Tree.

Altre metriche di valutazione utilizzate sono state gli score di *Accuracy* ed *F1*, controllati con il metodo cross-validation di Scikit learn. I risultati sono leggermente inferiori, ma si mantengono soddisfacenti.

```
Accuracy: 0.7911 (+/- 0.03)
F1-score: 0.7015 (+/- 0.05)
```

La *TrainingCurve* ci ha mostrato il variare dell'accuratezza all'aumentare della training size, dandoci una conferma generale della crescita di accuracy del modello all'aumentare dei dati.

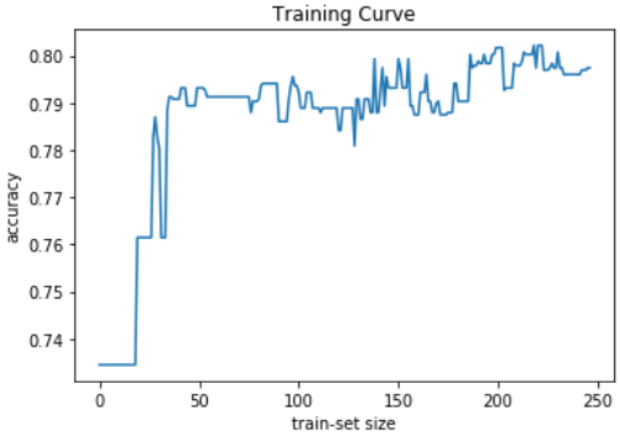

Fig.27 Training curve.

3.2 Classificazione con Random Forest

Per completezza abbiamo deciso di utilizzare un secondo algoritmo di classificazione, il **Random Forest**. Abbiamo proceduto in modo analogo a quanto fatto per il DecisionTree Classifier, cercando di trovare il miglior stimatore da utilizzare. Come nel caso del singolo albero, la tecnica di cross-validation è stata applicata ad un set di classificatori, ottenuti dalla combinazione di una lista di possibili valori per gli iperparametri, grazie al RandomizedSearchCV di sklearn.model_selection. Il classificatore risultante è stato poi utilizzato per allenare un modello sul training set ed è stato applicato al test set. Le feature selezionate sono state le stesse scelte nel modello precedente.

Valutando il modello con le metriche di *Accuracy* ed *F1* notiamo però un miglioramento rispetto al metodo precedente, nel caso dell'accuracy di circa tre punti percentuali ed è possibile riscontrare il 10% in più di precisione sui positivi.

```
Train Accuracy 0.8133874239350912
Train F1-score [0.87833906 0.59965187]

Test Accuracy 0.8277330809275911
Test F1-score [0.88834356 0.62318841]
              precision    recall  f1-score   support

           0       0.85      0.93      0.89      1552
           1       0.74      0.54      0.62       561

    accuracy                           0.83      2113
   macro avg       0.80      0.73      0.76      2113
weighted avg       0.82      0.83      0.82      2113
```

Fig.28 Accuracy e F1 score.

È possibile notare il miglioramento del modello anche nella Confusion Matrix e nella ROC Curve, la quale presenta un paio di punti percentuali in più.

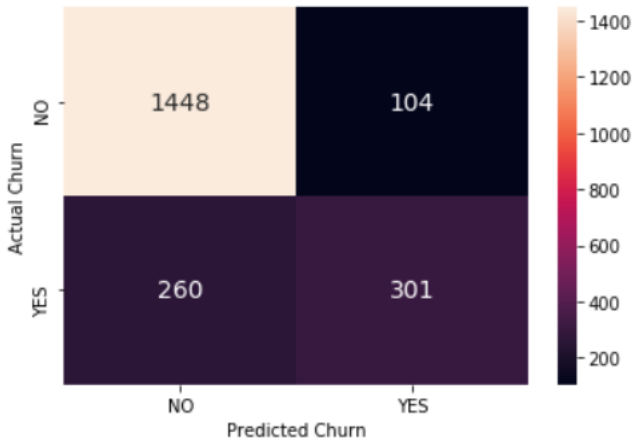

Fig.29 Confusion matrix per il Random Forest.

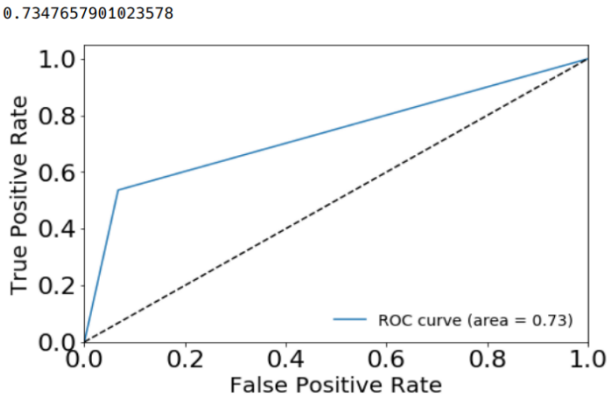

Fig.30 Curva ROC per il Random Forest.

Ulteriore conferma al miglioramento del modello, ci è stata data da un controllo sulle metriche di accuratezza grazie alla cross-validation, già applicata nel Decision Tree, e dalla Training Curve.

```
Accuracy: 0.8001 (+/- 0.03)
F1-score: 0.7140 (+/- 0.04)
```

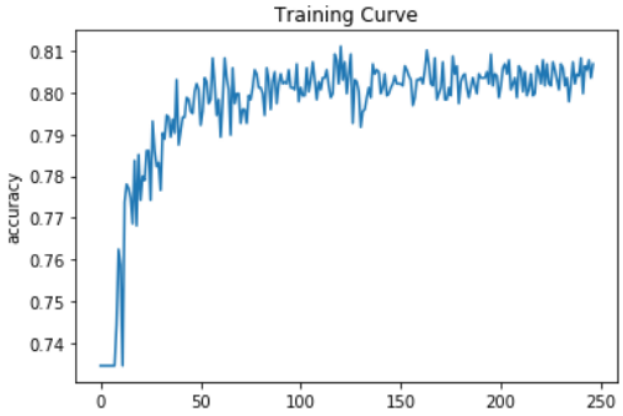

Fig.31 Training Curve per il Random Forest.

L'analisi ci porta dunque a confermare la maggior efficacia del Random Forest in scenari di questo tipo, dove il metodo d'ensemble riesce a performare con maggior accuratezza rispetto al singolo albero decisionale.

4. Pattern Mining

In preparazione all'attività di estrazione di pattern significativi dal nostro dataset, abbiamo prima di tutto importato, in aggiunta ai moduli con i quali abbiamo svolto le analisi precedenti, i moduli *apriori* e *operator*. Siamo poi passati alla preparazione del nostro dataset, assegnando ad ogni singolo valore delle feature un suffisso con il nome stesso della feature, così da distinguere i valori duplicati e quindi le feature a cui essi si riferiscono. Inoltre, per quanto riguarda i valori delle variabili numeriche, sono stati aggregati tramite la tecnica del *binning* (*Fig.32*).

StreamingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod	Churn	tenureBin	MonthlyChargesBin
No_StreamingTV	No_StreamingMovies	Month-to-month_Contract	Yes_PaperlessBilling	Electronic check_PaymentMethod	No_Churn	[0.0, 4.8)_tenureBin	[24.667, 31.333)_MonthlyChargesBin
No_StreamingTV	No_StreamingMovies	One year_Contract	No_PaperlessBilling	Mailed check_PaymentMethod	No_Churn	[33.6, 38.4)_tenureBin	[51.333, 58.0)_MonthlyChargesBin
No_StreamingTV	No_StreamingMovies	Month-to-month_Contract	Yes_PaperlessBilling	Mailed check_PaymentMethod	Yes_Churn	[0.0, 4.8)_tenureBin	[51.333, 58.0)_MonthlyChargesBin
No_StreamingTV	No_StreamingMovies	One year_Contract	No_PaperlessBilling	Bank transfer (automatic)_PaymentMethod	No_Churn	[43.2, 48.0)_tenureBin	[38.0, 44.667)_MonthlyChargesBin
No_StreamingTV	No_StreamingMovies	Month-to-month_Contract	Yes_PaperlessBilling	Electronic check_PaymentMethod	Yes_Churn	[0.0, 4.8)_tenureBin	[64.667, 71.333)_MonthlyChargesBin

Fig.32 Esempio di valori delle feature dopo la trasformazione.

In seguito ai primi test di estrazione di pattern, il primo aspetto che è emerso è stata la presenza nei risultati di pattern decisamente poco significativi. Ovvero, combinazioni di valori prive di informazioni effettivamente utili (e.g. 'Yes_Partner', 'No_PhoneService', 'No_DeviceProtection', 'Yes_PaperlessBilling'). Questi primi risultati sono dovuti alla diversa natura delle feature del dataset. Sono presenti, infatti, come già accennato sopra, feature relative alle informazioni personali dei clienti, altre relative ai servizi acquistati e altre ancora relative al tipo di contratto e pagamenti. Pertanto, l'intuizione che abbiamo avuto è stata quella di estrarre, dall'universo dei pattern e regole associative generate, dei sottoinsiemi di questi applicando dei filtri, così da ottenere delle combinazioni quanto più significative possibile.

Ad esempio, nei pattern in cui compaiono valori di feature relative ai servizi, è indispensabile, ai fini dell'indagine che stiamo portando avanti, poter discriminare se quel set di clienti ha effettivamente acquistato il servizio internet o meno. La regola zero che abbiamo applicato è stata quella di estrarre pattern e rule per categorie di feature. Ovvero, applicare la funzione di estrazione a dataframe contenenti solo un sottoinsieme di tutte le colonne del dataset di partenza. I sottoinsiemi coincidono appunto con le categorie di feature sopra citate (1. Informazioni riguardanti il cliente, 2. Informazioni su contratto e pagamenti, 3. Informazioni riguardanti i servizi offerti) (*Fig. 33*).

```
customer_info = ['gender', 'SeniorCitizen', 'Partner', 'Dependents']
contract_info = ['Contract', 'PaperlessBilling', 'PaymentMethod']
services = ['InternetService', 'MultipleLines', 'OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport',
            'StreamingTV', 'StreamingMovies']

n_top = 5 # Top 5
```

Fig.33 Tre gruppi gruppi di feature in analisi.

Da notare che l'unico caso in cui abbiamo pensato potesse aver più senso estrarre delle regole associative è sui servizi acquistati dal cliente, che coincide, almeno in linea teorica, con uno dei casi più esemplari di pattern mining, il carrello della spesa. Per questo, è possibile leggere le regole estratte allo stesso modo. In quasi la totalità dei casi abbiamo applicato un supporto minimo pari al 10% e un numero minimo di feature pari a quattro. Solo in alcuni casi di estrazione di pattern abbiamo ridotto il numero minimo delle feature a due e tre, laddove il numero delle feature coinvolte era minore o uguale a quattro. Per quanto riguarda le regole associative, abbiamo invece impostato un valore minimo di confidence pari al 60%. Abbiamo deciso di stabilire queste soglie in modo da provare ad estrarre qualcosa di interessante anche per valori bassi dei suddetti parametri.

4.1 Pattern e regole associative per categorie di feature

4.1.1 Informazioni riguardanti il cliente (supp_min = 10, zmin = 3)

Pattern con z=3 (Top 5)

	gender	SeniorCitizen	Partner	Dependents	support (%)
0		No	No	No	38.605708
1	Male	No		No	27.658668
2	Female	No		No	27.459889
3		No	Yes	Yes	23.654693
4	Female		No	No	23.498509

4.1.2 Informazioni su contratto e pagamenti (supp_min=10, zmin=2)

Patterns con z=2 (Top 5)

	Contract	PaperlessBilling	PaymentMethod	support (%)
0	Month-to-month	Yes		36.717308
1	Month-to-month		Electronic check	26.267216
2		Yes	Electronic check	24.733778
3	Month-to-month	No		18.301860
4		No	Mailed check	13.573761

4.1.3 Informazioni riguardanti i servizi offerti

4.1.3.1 Pattern

Giunti a questo punto dell'indagine, ci siamo orientati verso l'applicazione di un ulteriore filtro, andando a selezionare quei pattern che permettano di distinguere se il cliente ha acquistato il servizio internet o meno. Pertanto i risultati che andremo a presentare si distinguono in 'Internet service No' e 'Internet Service Yes'.

- *'Internet Service No'*

Per quanto riguarda i risultati che ricadono nella categoria di coloro che non hanno optato per il servizio internet, riteniamo sia interessante riportare il pattern più frequente e in nostra opinione più significativo che abbiamo estratto.

	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport	StreamingTV	StreamingMovies	support (%)
0	No	No internet service	No internet service	No internet service	No internet service	No internet service	No internet service	21.666903

- *'Internet Service Yes'*

Pattern con z=4 (Top 5)

	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport	StreamingTV	StreamingMovies	support (%)
0		Fiber optic	No	No		No			16.527048
1		Fiber optic	No		No	No			16.512850
2	Yes	Fiber optic	No			No			14.610251
3		Fiber optic		No	No	No			14.141701
4	Yes	Fiber optic					Yes	Yes	13.644754

4.1.3.2 Regole associative

Per quanto riguarda l'estrazione di regole, come già anticipato, abbiamo deciso di applicare il metodo di estrazione solamente per i servizi offerti dalla compagnia. Anche in questo caso abbiamo applicato la discriminazione 'Internet Yes'/'Internet No'. Abbiamo, inoltre, applicato un filtro che generasse dei sottoinsiemi quanto più rappresentativi possibile. In questo modo, abbiamo ottenuto due sottoinsiemi di regole: da una parte, le più 'forti', quelle cioè che presentano il valore *lift* più alto; dall'altra, le più 'popolari', le quali presentano un valore di *support* più alto.

- *'Internet Service Yes'*

Regole con lift maggiore (minimo 2)

	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport	StreamingTV	StreamingMovies	lift	confidence (%)	support (%)
0			Yes		Yes	<< Yes >>		Yes	2.407330	69.864865	10.506886
1			<< Yes >>	Yes	Yes	Yes			2.258305	64.738292	10.308107
2	Yes	Fiber optic			Yes		<< Yes >>	Yes	2.231617	85.772914	10.379100
3	Yes	Fiber optic			Yes		Yes	<< Yes >>	2.199162	85.306122	10.435894
4	Yes			Yes	<< Yes >>		Yes	Yes	2.179923	74.964838	10.095130
5		DSL			No		No	<< No >>	2.084262	82.417582	14.212693
6		DSL				No	<< No >>	No	2.071651	82.654249	12.196507
7	Yes	<< Fiber optic >>				No	Yes		2.032919	89.364162	12.281698

Regole con support maggiore

	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport	StreamingTV	StreamingMovies	lift	confidence (%)	support (%)
0		Fiber optic	No	<< No >>		No			1.50	65.95	25.06
1		Fiber optic	No		<< No >>	No			1.50	65.89	25.06
2		<< Fiber optic >>	No		No	No			1.52	66.92	24.68
3		Fiber optic	<< No >>		No	No			1.65	81.96	20.15
4		Fiber optic	No	No		<< No >>			1.68	82.97	19.92
5	Yes	Fiber optic	No					<< Yes >>	1.59	61.57	19.21
6	Yes	Fiber optic	No				<< Yes >>		1.58	60.75	19.21
7		Fiber optic			<< Yes >>		Yes	Yes	1.78	61.15	18.60
8	<< Yes >>	Fiber optic					Yes	Yes	1.74	73.36	18.60
9		Fiber optic	No		No	No	<< No >>		1.50	60.02	16.51
10		Fiber optic				No	No	<< No >>	1.75	69.38	15.68
11	Yes			<< Yes >>	Yes			Yes	1.85	63.72	14.13
12			Yes			<< Yes >>	Yes	Yes	2.26	65.68	11.00
13			<< Yes >>	Yes	Yes	Yes			2.26	64.74	10.31
14		<< DSL >>	Yes		Yes	Yes			1.82	62.69	10.12

- *'Internet Service No'*

Per tutti i clienti che non usufruiscono del servizio Internet, i valori delle feature relative ai servizi extra riportano come valore 'No'. Non ha senso, dunque, ricercare delle regole associative in merito.

4.2 Pattern e regole associative utilizzando tutte le feature

Gli stessi criteri di estrazione di pattern e di regole applicati sopra per le categorie di feature ('Internet Service Yes'/'Internet Service No'), sono stati in seguito applicati all'intero dataset, includendo così tutte le colonne che avevamo a disposizione. Abbiamo così ricavato le regole con lift più alto e con più alto supporto per l'intero dataset.

4.2.1 Pattern

- *'Internet Service No'*

Pattern più frequente

	PhoneService	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport	StreamingTV	StreamingMovies	support (%)
0	Yes	No	No internet service	No internet service	No internet service	No internet service	No internet service	No internet service	21.666903

- *'Internet Service Yes'*

Pattern con z=4 (Top 5)

	Dependents	PhoneService	InternetService	OnlineSecurity	TechSupport	Contract	PaperlessBilling	support (%)
0	No	Yes	Fiber optic				Yes	27.119125
1	No	Yes	Fiber optic	No				26.011643
2	No	Yes	Fiber optic		No			25.756070
3		Yes	Fiber optic		No	Month-to-month		25.500497
4		Yes	Fiber optic	No			Yes	25.429504

4.2.2 Regole Associative

Regole associative con lift maggiore (minimo 2)

Vedi Appendice (*Fig. 34*)

Regole associative con support maggiore (Top 20)

Vedi Appendice (*Fig. 35*)

4.3 Estrazione dei pattern per 'Churn Yes'

Infine, dopo l'esplorazione panoramica di cui sopra, abbiamo nuovamente spostato il focus di questa analisi sul target della nostra indagine, i clienti che hanno scelto di rescindere il contratto, ovvero, i clienti che ricadono nella categoria 'Churn=Yes'. La focalizzazione dell'analisi su questo relativamente piccolo sottoinsieme del dataset originario permette di poter fare delle considerazioni molto più significative di quelle che si potrebbero fare con le stesse analisi applicate sull'intero data set. Per dare un esempio concreto, riteniamo sia molto significativo riportare che il 55% dei clienti che hanno scelto di rescindere il contratto con la compagnia ricadano nel pattern [Yes_PhoneService, Fiber optic_InternetService, No_TechSupport, Month-to-month_Contract]. Riteniamo meno significativo riportare che gli stessi clienti rappresentano solamente il 4% del dataset globale.

	Partner	Dependents	Phone Service	Internet Service	Online Security	Online Backup	Device Protection	Tech Support	Contract	Paperless Billing	Payment Method	support(%)
0			Yes	Fiber optic				No	Month-to-month			55.27
1			Yes	Fiber optic	No			No	Month-to-month			49.49
2		No	Yes	Fiber optic	No			No	Month-to-month			42.85
3		No	Yes	Fiber optic	No			No	Month-to-month	Yes		36.11
4		No	Yes	Fiber optic	No			No	Month-to-month	Yes	Electronic check	26.69
5		No	Yes	Fiber optic	No	No	No	No	Month-to-month	Yes		20.49
6		No	Yes	Fiber optic	No	No	No	No	Month-to-month	Yes	Electronic check	15.46
7	No	No	Yes	Fiber optic	No	No	No	No	Month-to-month	Yes	Electronic check	11.77

Invece, per quanto riguarda le regole associative in questa parte, considerata la natura diversa delle feature, riteniamo che conoscere soltanto i pattern più significativi risulta già piuttosto efficace a profilare i clienti 'Yes_Churn'.

Conclusioni

In conclusione, le varie tecniche di analisi che abbiamo applicato al dataset hanno fatto emergere delle caratteristiche abbastanza chiare riguardanti gli utenti che decidono di rescindere il contratto.

Nella fase di Data Understanding, abbiamo visto che la maggior parte degli abbandoni si concentrava nei primi mesi dall'attivazione dei contratti ('tenure') e laddove il canone mensile era più alto rispetto alla media. Inoltre, è emersa in questa fase preliminare la tendenza a non usufruire dei servizi extra offerti dalla compagnia ('TechSupport', 'OnlineSecurity' etc.) dei clienti che sceglievano di rescindere il contratto dopo poco. Inoltre, un altro schema ricorrente si è rivelato essere la scelta della fibra ottica fra i servizi acquistati e la modalità di rinnovo del contratto mensile.

Successivamente, durante le operazioni di clustering, abbiamo riscontrato chiaramente come quei clienti che scelgono di abbandonare la compagnia rappresentino un gruppo distinguibile grazie alle caratteristiche di cui sopra. In particolare, mettendo in relazione i cluster ricavati mediante l'algoritmo di Kmodes con la feature 'churn', è emerso come un sottoinsieme considerevole dei clienti che abbandonano appartenga al cluster caratterizzato dagli stessi valori delle feature individuati nella fase di Data Understanding.

Nella successiva fase di classificazione, i due modelli utilizzati, DecisionTree e RandomForest, hanno riportato entrambi dei risultati interessanti da diversi punti di vista. Il primo, mediante l'indice di importanza delle feature, ha confermato le feature significative relative al 'churn' già individuate nelle fasi precedenti. Il secondo invece, grazie alle sue potenzialità, ha permesso di raggiungere un buon livello di accuratezza – oltre l'80% – della classificazione. Purtroppo però,

non altrettanto soddisfacenti sono stati i valori di precision e recall dei positivi (churn yes) che si attestano rispettivamente al 74 e al 54 %.

Infine, nella sezione dedicata al Pattern Mining e, nello specifico, la parte relativa al sottoinsieme del dataset corrispondente a ‘churn yes’ è facilmente riscontrabile come il pattern con maggiore supporto corrisponda ai risultati emersi nelle fasi precedenti. Riteniamo quindi, con ragionevole certezza, di poter formulare il profilo di una significativa percentuale dei clienti che scelgono di abbandonare la compagnia: persone che acquistano il servizio internet in modalità fibra ottica, scartano tutti i servizi internet accessori (quali supporto, protezione dei dispositivi, backup e sicurezza) e che scelgono la formula di contratto con rinnovo mensile.

Si potrebbe quindi consigliare alla compagnia una revisione del servizio della fibra ottica, ad esempio, valutare se funzioni adeguatamente e se risponda alle aspettative dei clienti o se sia un suo malfunzionamento ad indurre clienti a cambiare fornitore. Si può altresì intuire come sia più facile rescindere dal contratto se il vincolo è mensile, invece che annuale. Agevolazioni per contratti più lunghi potrebbero quindi facilitare la retention della clientela. I servizi accessori ‘OnlineSecurity’, ‘OnlineBackup’, ‘TechSupport’, ‘DeviceProtection’ per cui si vede un pattern di abbandono per i clienti che non li attivano, potrebbero essere sentore di una tipologia di cliente meno incline ad approfondire le offerte della compagnia in termini di servizi Internet. La compagnia stessa potrebbe capire se offrendo questo tipo di extra, magari ad un prezzo agevolato, possa aumentare il grado di fedeltà e diminuire quindi il churn. Questi potrebbero essere i piani d’azione che ci troveremmo a consigliare alla compagnia per diminuire la churn rate.

Appendice

	SeniorCitizen	Partner	Dependents	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport	StreamingTV	Streaming Movies	Contract	PaperlessBilling	PaymentMethod	Churn	MonthlyChargesBin	lift	confidence (%)	support (%)
0				Yes	No	No	No internet service	No internet service	No internet service	No internet service	No internet service	No internet service					<< [18.0, 24.667) >>	5.07	100.00	16.81
1				Yes		No	No internet service	No internet service	<< No internet service >>	No internet service	No internet service	No internet service						4.62	100.00	21.67
2						<< No >>								No		No	[18.0, 24.667)	4.57	99.00	12.79
3							Yes		Yes	Yes			<< Two year >>					2.79	67.04	10.12
4							Yes			<< Yes >>			Two year			No		2.77	80.28	10.29
5			No	Yes		Fiber optic	No			No			Month-to-month	Yes	Electronic check	<< Yes >>		2.62	69.50	10.19
6							<< Yes >>			Yes			Two year			No		2.56	73.48	11.25
7	No	Yes	<< Yes >>	Yes		No	No internet service	No internet service	No internet service	No internet service	No internet service	No internet service						2.46	73.56	10.10
8									<< Yes >>			Yes	Two year			No		2.43	83.61	10.39
9				Yes	Yes	Fiber optic			Yes		<< Yes >>	Yes						2.23	85.77	10.38
10				Yes	Yes	Fiber optic			Yes		Yes	<< Yes >>						2.20	85.31	10.44
11	No					<< DSL >>	Yes				No					No		2.20	75.48	10.31
12			No	Yes		Fiber optic				No			Month-to-month	Yes	<< Electronic check >>	Yes		2.17	72.75	10.58
13						DSL			No		<< No >>	No	Month-to-month					2.16	86.21	10.61
14						DSL			No		No	<< No >>	Month-to-month					2.16	85.30	10.72
15				Yes	Yes	<< Fiber optic >>	No			No			Month-to-month	Yes				2.13	93.64	10.93
16								<< Yes >>			Yes		Two year			No		2.11	72.82	10.24
17		No		Yes	<< No >>	No	No internet service	No internet service	No internet service	No internet service	No internet service	No internet service					[18.0, 24.667)	1.91	92.14	10.29
18						DSL			<< No >>		No	No	Month-to-month					1.91	83.75	10.92
19				Yes		Fiber optic	No		No	<< No >>	No		Month-to-month					1.87	92.42	10.31
20				Yes	<< Yes >>	Fiber optic			Yes		Yes	Yes						1.86	78.28	11.37

Fig.34 Regole associative con lift maggiore considerando tutte le feature.

	SeniorCitizen	Partner	Dependents	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod	Churn	MonthlyChargesBin	lift	confidence (%)	support (%)
0			No	Yes		<< Fiber optic >>								Yes				1.55	68.14	39.80
1			No	Yes		Fiber optic	<< No >>											1.52	75.27	34.56
2			No	Yes		Fiber optic				<< No >>								1.51	74.53	34.56
3			No	Yes	<< Yes >>	Fiber optic												1.47	62.12	34.56
4			No	Yes		Fiber optic							<< Month-to-month >>					1.33	73.17	34.56
5			No	Yes		Fiber optic								<< Yes >>				1.33	78.47	34.56
6		<< No >>	No	Yes		Fiber optic												1.20	61.91	34.56
7	<< No >>		No	Yes		Fiber optic												0.81	68.28	34.56
8			<< No >>	Yes		Fiber optic								Yes				1.14	79.75	34.01
9	No			Yes		Fiber optic										<< No >>		0.82	60.09	32.16
10				Yes		Fiber optic	No	<< No >>										1.42	62.16	32.05
11				Yes		Fiber optic	No		<< No >>									1.39	60.92	32.05
12				Yes		Fiber optic							Month-to-month		<< Electronic check >>			1.83	61.42	30.21
13				Yes	Yes	Fiber optic					<< Yes >>							1.65	63.26	27.52
14				Yes	Yes	Fiber optic						<< Yes >>						1.63	63.42	27.52
15			No	<< Yes >>		Fiber optic								Yes				1.11	100.00	27.12
16				Yes					<< Yes >>		Yes	Yes						1.87	64.14	24.71
17		<< Yes >>		Yes							Yes					No		1.29	62.38	23.74
18	No					<< DSL >>					No	No						1.75	60.16	23.70
19	No			Yes		DSL					<< No >>							1.53	61.11	22.49
20	No			Yes		DSL						<< No >>						1.53	60.48	22.49
21	No			Yes	<< No >>	DSL												1.27	61.24	22.49

Fig.35 Regole associative con supporto maggiore considerando tutte le feature.