Studenti:

Dell'Olio Domenico (mat. 699781)
Delvecchio Giovanni Pio (mat. 697165)
Disabato Raffaele (mat. 708797)
Lamantea Giuseppe (mat. 706593)

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

Relazione Caso di Studio - Ingegneria della Conoscenza

Applicazione di tecniche di apprendimento supervisionato, clustering e dell'apprendimento di una rete Bayesiana per l'analisi nel dominio turistico-alberghiero

Repository: https://github.com/llamandtea/icon2021

Obiettivi e ispirazione

L'obiettivo principale del caso di studio è stato estrarre conoscenza in merito ai fattori che comportano la cancellazione di una prenotazione per una struttura alberghiera.

Considerando il paper:

Predicting Hotel Bookings Cancellation with a Machine Learning Classification Model - IEEE Conference Publication

abbiamo lavorato con l'obiettivo di migliorare i risultati, anche utilizzando metodi più semplici.

E' stata posta anche la problematica di predire quanti giorni prima, rispetto all'arrivo dichiarato, viene cancellata una prenotazione.

Sono state attuate anche alcune strategie di Clustering per poter ottenere automaticamente dei gruppi di prenotazioni correlate tra loro e sono state costruite delle reti Bayesiane per estrarre correlazioni tra features, oltre che verificare le prestazioni del classificatore Bayesiano associato, rispetto al problema di individuare possibili cancellazioni.

Strumenti adottati

A livello implementativo, è stato principalmente utilizzato il linguaggio **Python**, per la flessibilità di utilizzo e la grande quantità di librerie disponibili per task di apprendimento automatico.
La gestione del dataset è stata effettuata principalmente tramite la libreria **Pandas**, mentre la costruzione di modelli di apprendimento (sia supervisionato che non supervisionato) è stata effettuata tramite le librerie **scikit-learn** e **kmodes**. Infine, il modello di apprendimento probabilistico è stato costruito tramite la piattaforma **Weka**.

Parte 0: Descrizione dei dataset e preprocessing

Sono stati utilizzati due datasets separati, riferiti a tipi di strutture differenti:

- un hotel resort;
- un hotel di città;

Entrambi gli hotel hanno un giudizio di 4 stelle e sono rimasti anonimi, come anonime sono le varie prenotazioni. Rispettivamente, i due datasets contengono 40,060 e 79,330 esempi, suddivisi in 31 features, contenenti informazioni su prenotazioni avvenute tra 1 Luglio 2015 e 31 Agosto 2017.

E' possibile visionare la descrizione completa di entrambi i datasets al link:

https://www.sciencedirect.com/science/article/pii/S2352340918315191

Considerando che il dataset contiene sia attributi categorici che continui, è stato effettuato un lavoro di preprocessing preliminare analogo a quello effettuato per il paper sopra riportato, a cui si sono susseguite diverse finiture ed adattamenti a seconda del metodo specifico da applicare.

In particolare il preprocessing preliminare e comune alle varie tecniche ha comportato le seguenti modifiche:

- creazione di una nuova feature "Season" che mappa le date di arrivo nella stagione corrispondente, considerando che le prenotazioni avvengono tutte nell'emisfero boreale;
- per le prenotazioni cancellate, il valore della feature "Lead Time" è stato mappato come la distanza, in numero di giorni, fra data di prenotazione e data di cancellazione. Per le altre, è rimasta la distanza in giorni tra la registrazione della prenotazione e la data di arrivo;
- è stata aggiunta una feature relativa all' "ADR" (Average Daily Rate) medio calcolata in base ai gruppi unici ottenuti attraverso le combinazioni dei valori delle features relative a stanza prenotata, settimana dell'anno di arrivo, canale di distribuzione, e anno di arrivo;
- sono state raggruppate le features "Children" e "Babies" in "Minors", ottenendo il numero di minori totali considerati all'interno della prenotazione;
- sono state raggruppate le features "Stays In Weekend Nights" e "Stays In Week Nights" in "Staying", ottenendo il numero totale di giorni di permanenza all'interno della struttura;
- è stata creata la feature "Cancel Rate", che indica il rateo di prenotazioni cancellate, rispetto alle prenotazioni totali effettuate da un utente in passato. In particolare, questa feature è calcolata come "Previous Cancellations"/("Previous Cancellations"+"Previous Bookings Not Canceled");
- è stata convertita la feature "Days In Waiting List" nella feature booleana "Was In Waiting List";
- sono state rimosse le colonne considerate poco rilevanti allo scopo prefissato o fuorvianti (vedasi la caratteristica relativa al paese che in caso di cancellazione, offriva sempre il valore 'PRT' di default).

Parte 1: Replicare ed Estendere

Per poter predire l'eventuale cancellazione della prenotazione è stata utilizzata la tecnica Gradient Boosting, una tecnica di ensemble learning con alberi di decisione che si è mostrata la modalità più valida tra quelle testate (Random Forest ed Adaboost).

Si riportano i fondamenti teorici di Gradient Boosting:

vengono creati una serie di "Weak learners" di classe H, ovvero classificatori più o meno correlati al valore effettivo da predire; in questo caso alberi di decisione molto semplici. Ognuno di essi ha un peso associato, che ad ogni passo si calibra in modo da minimizzare la funzione di log loss:

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)).$$

Dove:

- γ_m è il peso associato all' m-esimo albero costruito;
- L è la funzione di log loss;
- y_i è il valore da stimare;
- F_{m-1} è lo stimatore ottenuto dal passo precedente;
- x_i è l'esempio corrente;
- γ è il peso per l'albero considerato ad ogni passo della ricerca dell'argmin;
- h_m è l'albero corrente.

Questa tecnica è stata testata attraverso una *k-fold cross validation.* Essa consiste nella suddivisione del training set in k partizioni, di cui, iterativamente, si utilizzano k-1 partizioni per l'addestramento del modello e la restante viene utilizzata per testarlo.

In questo modo è possibile valutare la bontà del modello nel caso in cui non siano disponibili nuovi dati di test, considerando la media delle metriche ottenute durante le varie iterazioni.

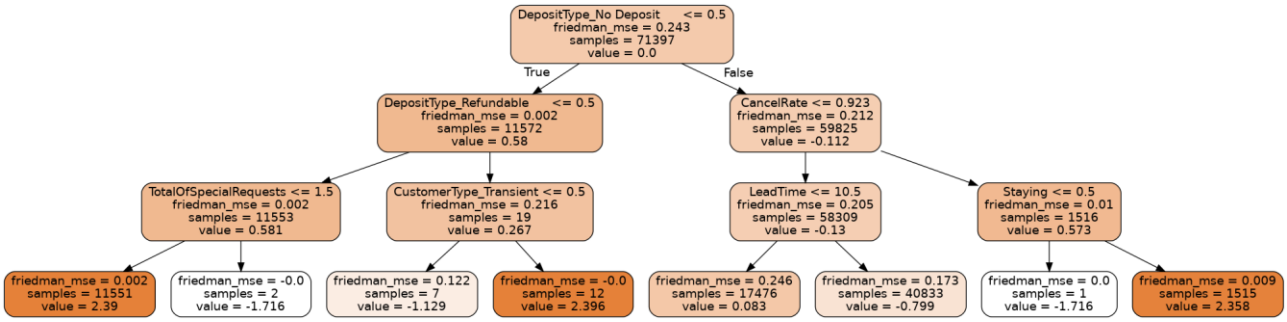
Applicando Gradient Boosting con 10-fold cross validation al dataset preprocessato sugli hotel in città si hanno le seguenti metriche:

Mean score	0.8249842430354215
Mean precision	0.8454021932577733
Mean recall	0.8008877139551919
Mean f-score (beta = 2.0)	0.8019628243143375

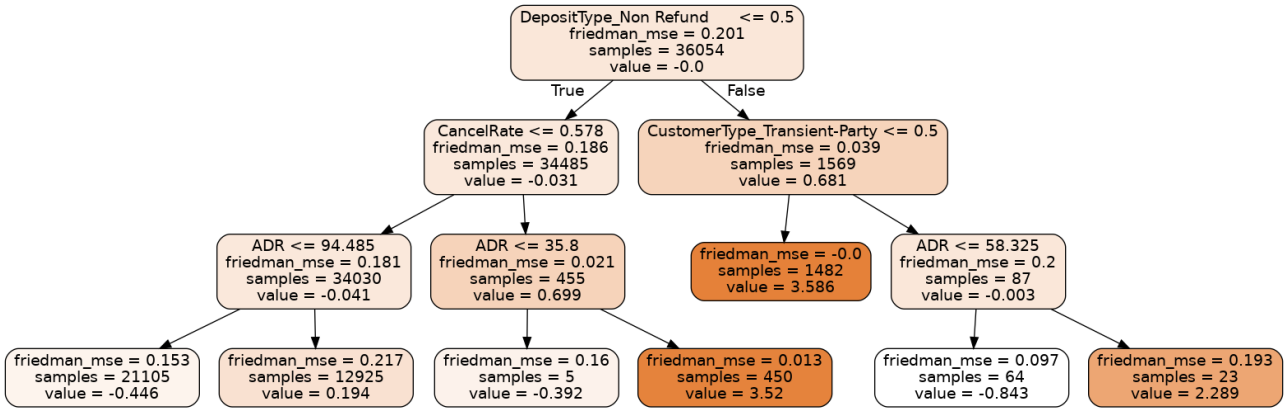
Analogamente, per il dataset preprocessato sui resort:

Mean score	0.8235147279081378
Mean precision	0.8399238712482248
Mean recall	0.7033261928477874
Mean f-score (beta = 2.0)	0.7093109808789742

Di seguito riportato l'albero di decisione migliore ottenuto con Gradient Boosting per gli hotel in città:



e l'albero di decisione migliore ottenuto per i resort:



I valori negativi ottenuti sulle foglie sono assimilabili a 0 e sono ottenuti per via di come opera inerentemente l'algoritmo.

Si osserva che, nonostante un minor numero di feature avrebbe ridotto il sovradattamento dei dati, le metriche ottenute si avvicinano a quelle del paper considerato, senza superarlo, ma ottenendo comunque buoni risultati.

Di seguito si riporta la tabella con gli score degli alberi costruiti dagli autori del paper (per confronto):

TABLE I. PERFORMANCE METRICS FOR THE 1ST JULY 2017

Hotel.	*Dataset*	*Acc.*	*AUC*	*Prec.*	*Sensit.*	*Specif.*
H1	Train	0.846	0.910	0.839	0.626	0.950
	Test	0.842	0.877	0.811	0.603	0.941
H2	Train	0.857	0.934	0.876	0.793	0.909
	Test	0.849	0.922	0.869	0.779	0.905

Dove H1 è il dataset sul resort ed H2 è il dataset sull'hotel di città.

Per il problema della previsione del preavviso di cancellazione è stato scelto di applicare un semplice albero di regressione, in quanto permetteva di eseguire una previsione migliore di quella ottenuta con Gradient Boosting Regression (di sklearn) e di gestire la natura ibrida del dataset (feature miste).

Un albero di regressione è una variante di un albero di decisione che permette predizioni su features continue, come, in questo caso, il numero di giorni di preavviso di cancellazione.

E' ottenuto eseguendo degli split sul dataset, scelti in base alle feature che massimizzano l'information gain.

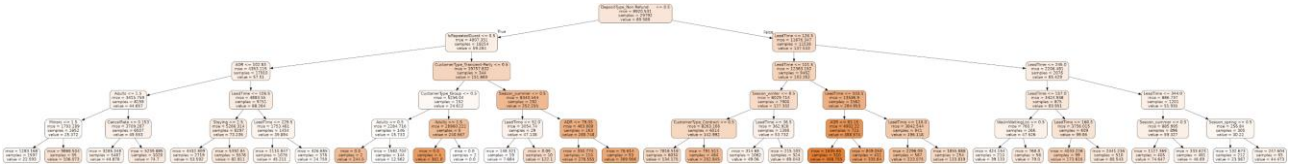
Si sono ottenute le seguente metriche per l'albero di regressione (per la predizione della feature "Cancellation Minus Arrival", ovvero la feature ingegnerizzata per indicare il numero di giorni di preavviso) costruito sul dataset sugli hotel in città:

Mean score	0.5373844930233377
Mean Squared Error	4098.731652535334
Mean Absolute Error	46.8624307598206
Max Error	314.5971449058219

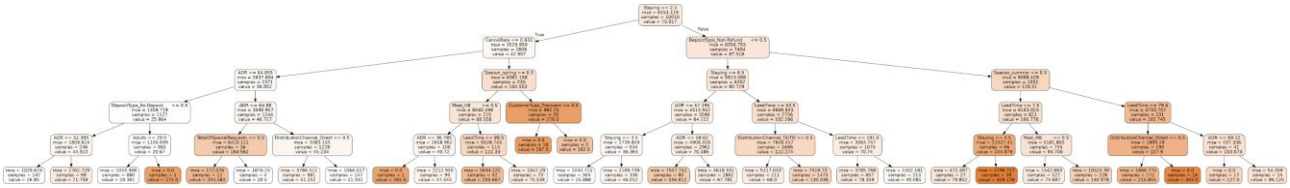
Analogamente per il dataset sul resort:

Mean score	0.3359529553548374
Mean Squared Error	4103.445093433407
Mean Absolute Error	47.16910332774184
Max Error	292.7863924343863

Di seguito riportato l'albero di regressione ottenuto per il dataset sugli hotel in città:



e per il dataset sui resort:



Per la costruzione di entrambi i modelli, sono stati eseguiti diversi test per determinare i parametri più adatti all'apprendimento dei classificatori e dei regressori. Tra questi, la variabile più importante è stata **max_depth**, che impone una profondità massima durante la costruzione degli alberi. Si è cercato di mantenere il valore di tale parametro basso in maniera tale da evitare situazioni di *overfitting*. In particolare, nella costruzione dell'albero di regressione, è stata scelta una profondità massima pari a 6, valore che è risultato un giusto compromesso tra accuratezza e generalità.

Parte 2: Clustering

Al fine di individuare delle classi di clientela che permettano di scoprire pattern nei valori delle feature utilizzate dal dataset, è stata utilizzata una tecnica di apprendimento non supervisionato, ovvero il Clustering. Infatti, il fine di tale tecnica è di individuare dei **cluster**, ovvero gruppi di esempi più simili a dei centroidi calcolati automaticamente.
Il clustering può essere di due tipi:

- **hard clustering**, che prevede un'assegnazione statica di ogni esempio ad una classe di appartenenza;
- **soft cluster**, che adotta delle distribuzione di probabilità sulle classi associate ad ogni esempio.

Nel caso di studio è stata adottata la prima tipologia di clustering. Al momento, le principali tecniche di hard clustering adottate sono **k-means** e **k-modes**: l'adozione di un algoritmo rispetto ad un altro dipende dalla natura del dataset. Infatti, l'algoritmo K-Means è formulato per dataset contenenti solamente features continue, mentre K-Modes è formulato per dataset contenenti solamente features categoriche.
La principale problematica incontrata durante l'applicazione del clustering sul dataset è la sua natura ibrida: è stato infatti necessario rifattorizzare il dataset in base all'approccio scelto.

Inizialmente è stato utilizzato l'algoritmo K-Means, che ha richiesto la conversione delle features utilizzate in feature continue e normalizzate, tramite una prima discretizzazione di tutte le variabili, ed in seguito la loro suddivisione in variabili indicatrici binarie. Tuttavia l'adozione di tale modello si è presto rivelata abbastanza macchinosa e produceva risultati logicamente poco significativi.
Dopo un approfondimento sugli algoritmi di clustering disponibili, è stato scelto di utilizzare l'algoritmo K-Modes, fermandosi alla sola discretizzazione delle variabili del dataset nel processo di rifattorizzazione. Tale discretizzazione è stata eseguita, quando possibile, utilizzando degli intervalli (*bin*) significativi (es per "Lead Time"), altrimenti in maniera automatica. È stata esplorata anche la possibilità di utilizzare l'algoritmo K-Prototypes per la clusterizzazione di dataset ibridi, ma le sue prestazioni rilevate sono peggiori rispetto alle tecniche precedentemente menzionate

I cluster di clienti individuati sono i seguenti:

IsCanceled	Meal	Distribution Channel	Is Repeated Guest	Deposit Type	Customer Type	Was Waiting List	Season	ADR	Lead Time	Booking Changes	Adults	Cancel Rate	Minors	Staying	Total Of Special Requests	N. Examples
No	BB	TA/TO	No	No Deposit	Transient	No	winter	(0, 103.593]	(0, 7.0]	(0, 5.0]	(1.0, 4.0]	(0, 0.5]	(0, 2.0]	(0, 3.0]	(0, 1.0]	20260
No	BB	TA/TO	No	No Deposit	Transient	No	spring	(103.593, 207.187]	(7.0, 30.0]	(0, 5.0]	(1.0, 4.0]	(0, 0.5]	(0, 2.0]	(0, 3.0]	(0, 1.0]	18550
No	BB	TA/TO	No	No Deposit	Transient-Party	No	autumn	(0, 103.593]	(30.0, 90.0]	(0, 5.0]	(1.0, 4.0]	(0, 0.5]	(0, 2.0]	(0, 3.0]	(0, 1.0]	10547
No	BB	TA/TO	No	No Deposit	Transient	No	summer	(0, 103.593]	(30.0, 90.0]	(0. 5.0]	(1.0, 4.0]	(0, 0.5]	(0, 2.0]	(3.0, 7.0]	(0, 1.0]	9977
Yes	BB	TA/TO	No	Non Refund	Transient	No	summer	(0, 103.593]	(30.0, 90.0]	(0, 5.0]	(1.0, 4.0]	(0, 0.5]	(0, 2.0]	(0, 3.0]	(0, 1.0]	10714
Yes	BB	TA/TO	No	No Deposit	Transient	No	spring	(103.593, 207.187]	(0, 7.0]	(0, 5.0]	(1.0, 4.0]	(0, 0.5]	(0, 2.0]	(0 3.0]	(0, 1.0]	9282

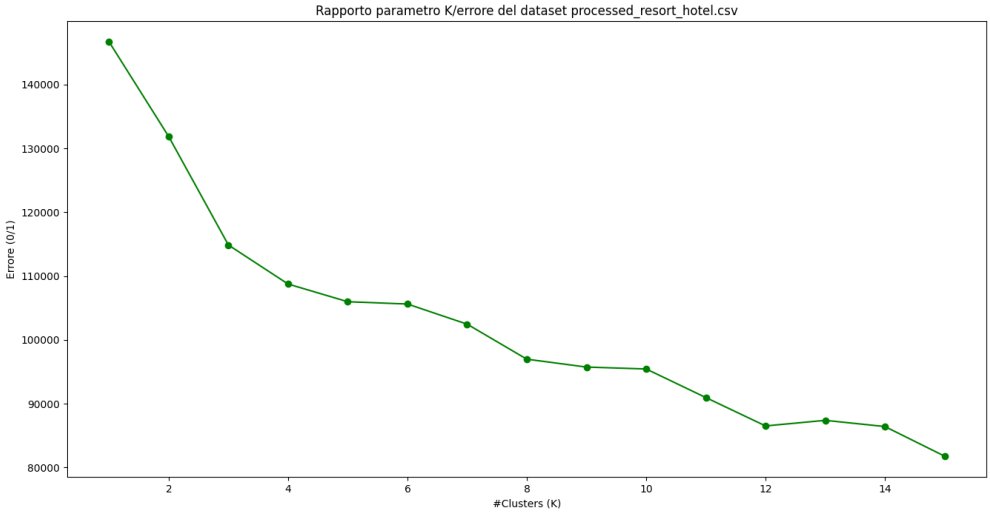
Cluster dei clienti degli hotel di città

IsCanceled	Meal	Distribution Channel	Is Repeated Guest	Deposit Type	Customer Type	Was Waiting List	Season	ADR	Lead Time	Booking Changes	Adults	Cancel Rate	Minors	Staying	Total Of Special Requests	N. Examples
No	BB	Direct	No	No Deposit	Transient	No	winter	(0, 115.21]	(0, 7.0]	(0, 5.0]	(0, 1.0]	(0, 0.5]	(0, 2.0]	(0, 3.0]	(0, 1.0]	7649
No	BB	TA/TO	No	No Deposit	Transient	No	summer	(115.21, 230.42]	(0, 7.0]	(0, 5.0]	(1.0, 55.0]	(0, 0.5]	(0, 2.0]	(3.0, 7.0]	(0, 1.0]	12170
No	BB	TA/TO	No	No Deposit	Transient	No	autumn	(0, 115.21]	(0, 7.0]	(0, 5.0]	(1.0, 55.0]	(0, 0.5]	(0, 2.0]	(0, 3.0]	(0, 1.0]	12219
No	BB	TA/TO	No	No Deposit	Transient	No	spring	(0, 115.21]	(30.0, 90.0]	(0, 5.0]	(1.0, 55.0]	(0, 0.5]	(0, 2.0]	(3.0, 7.0]	(0, 1.0]	8022

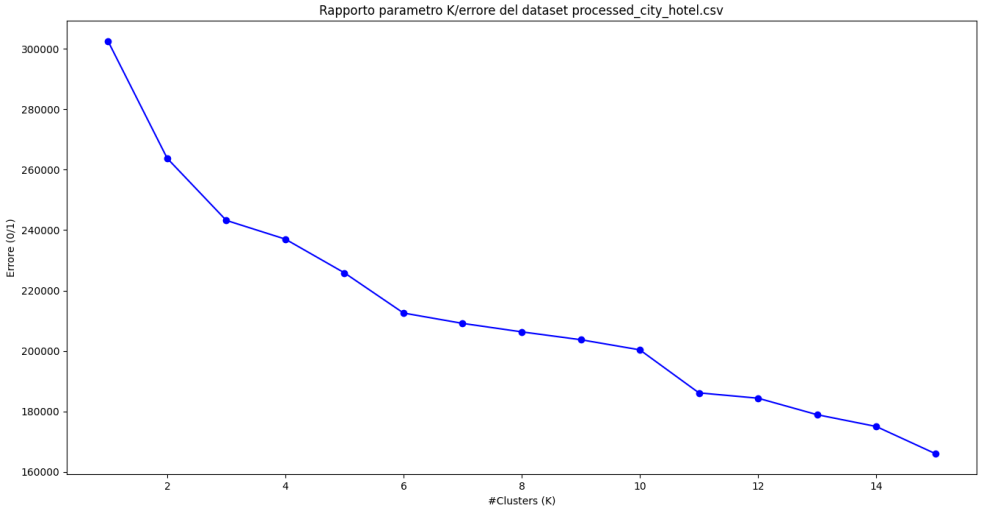
Cluster dei clienti degli hotel di resort

L’apprendimento dei cluster è stato eseguito attraverso ricerca locale con random restart su un totale di 50 iterazioni (considerata la minore efficienza dell’algoritmo rispetto al k-means).

L’algoritmo K-Modes, inoltre, prevede la scelta del parametro che indica il numero di centroidi da individuare: per tale selezione, è stato adottato il “metodo del gomito”. Esso permette di Individuare il K minimo che diminuisce significativamente l’errore associato al modello prima della sua stabilizzazione attraverso la valutazione del plot di un grafico errore/k e la selezione del K associato al gomito della curva.



Elbow plot per il dataset relativo agli hotel Resort



Elbow plot relativo agli hotel di città

Data la non-regolarità della curva, è stato scelto il primo punto dopo un cambio significativo di inclinazione e che individuasse un numero non troppo esteso di cluster (rispettivamente 4 e 6).

Parte 3: Belief Network

L'apprendimento automatico della Belief Network e delle relative probabilità associate è stato eseguito al fine di condensare gli scopi delle due parti precedenti, ovvero fornire un modello predittivo alternativo per la possibile cancellazione di una prenotazione e allo stesso tempo fornire una struttura attraverso la quale è forse possibile trarre informazioni non banali sulla correlazione delle feature ed eseguire altre interrogazioni di interesse.

Una rete bayesiana è un modello che rappresenta in forma grafica una distribuzione di probabilità congiunta su più variabili, alcune delle quali sono dipendenti le une dalle altre. Una rete è costituita, infatti, da un grafo orientato aciclico, che rappresenta la struttura delle relazioni di dipendenza tra le variabili nel modello, e da una serie di funzioni di distribuzione di probabilità condizionata e a priori che coinvolgono una o più variabili. In particolare, la rete rappresenta un ordinamento tra le features in termini di variabili "genitori" e variabili"figlie". Le prime sono l'insieme minimale dei predecessori delle seconde tale che gli altri predecessori delle figlie siano condizionatamente indipendenti da queste ultime data l'evidenza dei genitori. Questa relazione è espressa nella rete con la presenza di un arco direzionato da un nodo con etichetta di una variabile genitore verso uno con etichetta di una variabile dipendente da essa.

La struttura può essere nota a priori oppure deve essere appresa attraverso i dati.Il caso in esame si pone nella seconda classe di problemi, pertanto è necessario selezionare il migliore tra gli ordinamenti possibili sulle variabili. Individuare tale ordinamento implica una procedura di ricerca nello spazio dei possibili grafi aciclici orientati, con obiettivo la minimizzazione o la massimizzazione di una funzione di *score* calcolata su ciascuna struttura d'interesse. Inoltre è anche teoricamente possibile introdurre della conoscenza pregressa per poter ridurre lo spazio di ricerca, indicando ad esempio relazioni note tra le variabili. Tuttavia la piattaforma scelta, **Weka**, non permette l'aggiunta esplicita di tale conoscenza pregressa sulla struttura. Pertanto, sono stati eseguiti test preliminari al fine di ottenere una struttura che avesse una semantica il più possibile realistica, oltre che uno score ottimale a seguito di una valutazione **10-fold**. Inoltre tali test hanno portato anche alla selezione di due ulteriori parametri richiesti per la costruzione di una **"BayesNet"** con algoritmo di ricerca locale **"hill-climber"**, ovvero:

- *maxNrOfParents:* il numero massimo di genitori che un nodo possa avere. Esso influenza l'eventuale adattamento del modello ai dati e la complessità della rete - posto a **3**;
- *scoreType:* lo score da utilizzare per ottenere la rete dalla struttura migliore - selezionato l'**AIC:** Akaike Information Criterion, simile al BIC trattato in programma;

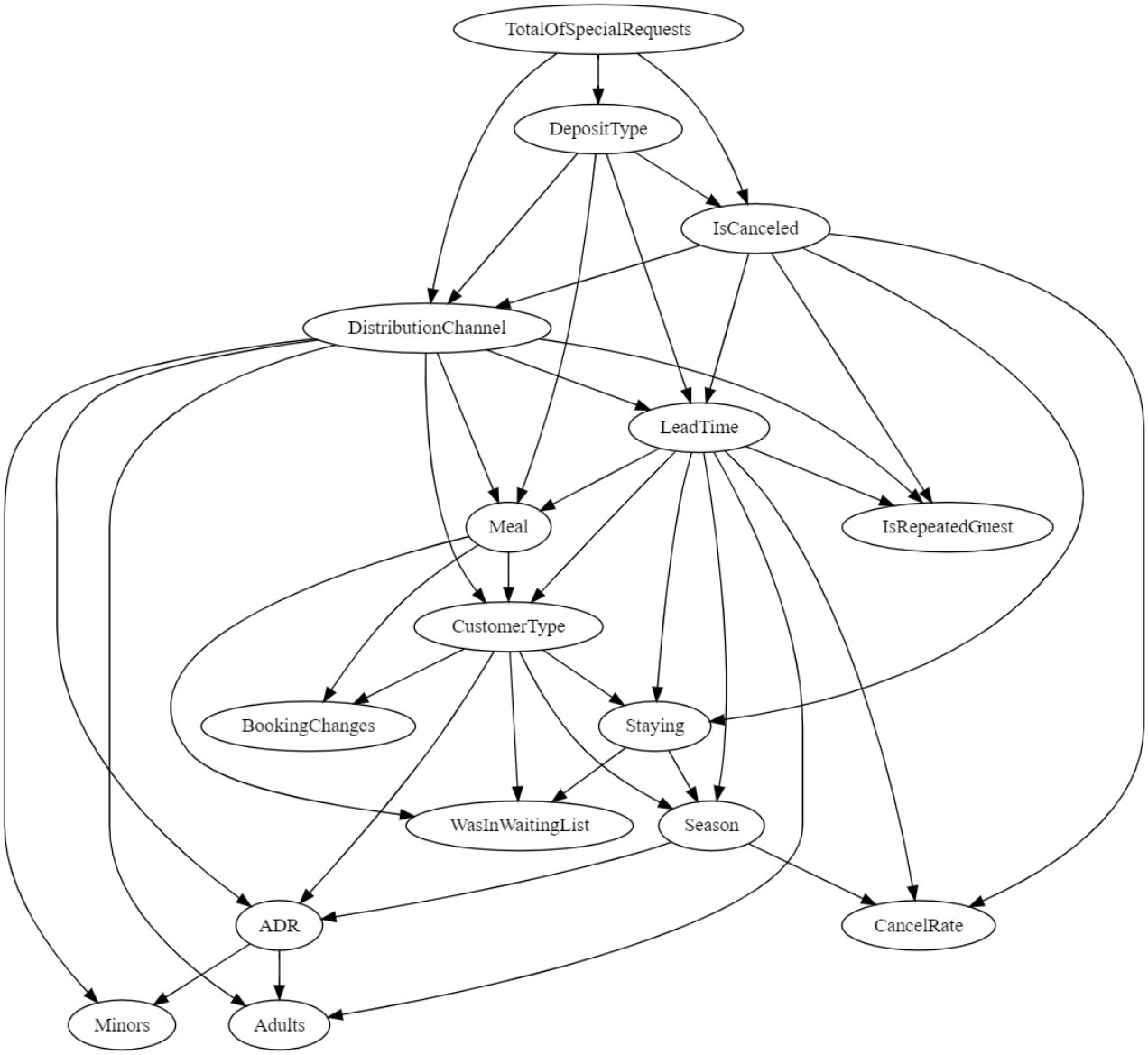
Inoltre è stato utilizzato lo stimatore "**SimpleEstimator**" per calcolare le probabilità a partire dai dati, applicando una stima campionaria con fattore di correzione *alpha* (a 0.5 di default - lasciato tale poiché non ci sono note le distribuzioni dei dati o eventuali pseudo conteggi) :

$$P(X = x) = \frac{\#esempi\ con\ X = x\ + \alpha}{\#esempi\ di\ training\ + \alpha}$$

$$P(X = x \mid parents(X) = j) = \frac{\#esempi\ con\ X = x\ e\ parents(X) = j\ + \alpha}{\#esempi\ con\ parents(X) = j\ + \alpha}$$

Presentiamo qui di seguito le reti costruite con relative prestazioni sui dataset di riferimento (la visualizzazione dei grafi è stata ottenuta attraverso questo sito). Esse sono prossime a quelle ottenute con gli alberi ma sicuramente di poco minori. Osserviamo quindi che probabilmente tali modelli potrebbero essere migliorati aggiungendo della conoscenza di fondo da parte di esperti per la costruzione della struttura o in generale di fornire una struttura predefinita.

Hotel Resort:



Rete Bayesiana costruita a partire dal dataset degli hotel Resort

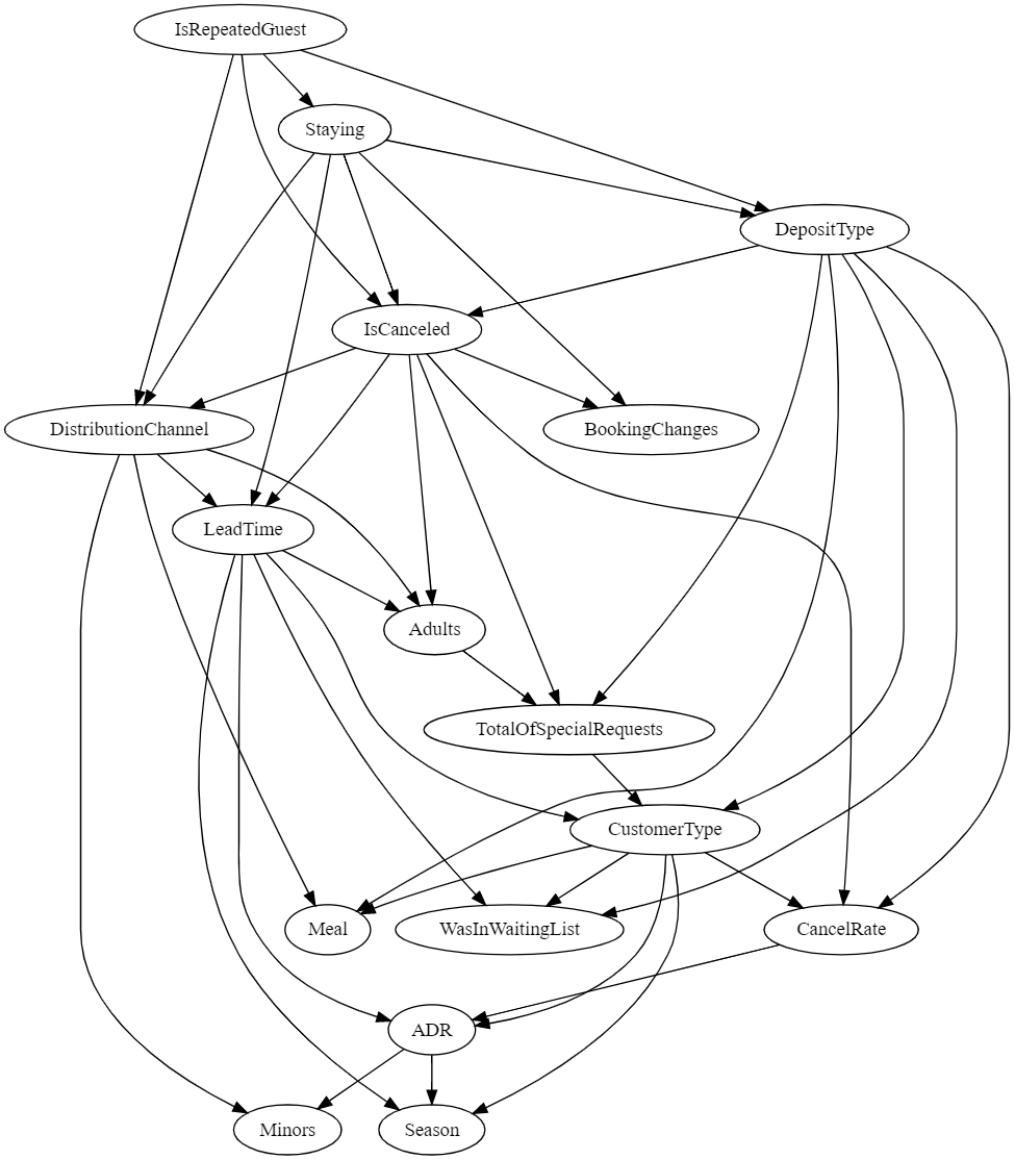
Qui di seguito si indicano i valori per la valutazione della struttura e le prestazioni del classificatore associato, fornite sempre da Weka:

AIC	-338124.93
Entropia	-336353.93
MDL	-345738.58

	Accuracy	**Precision**	**Recall**	**F-Measure**	**FP-Ratio**
Cancellato	80.48%	0.800	0.366	0.510	0.026
Non Cancellato		0.842	0.974	0.878	0.634
Media		0.811	0.805	0.776	0.465

Classificato come ↱	**Cancellato**	**Non Cancellato**
Cancellato	4069	7053
Non Cancellato	764	28174

Hotel di Città:



Rete Bayesiana costruita a partire dal dataset degli hotel di città

Qui di seguito si indicano i valori per la valutazione della struttura e le prestazioni del classificatore associato:

AIC	-660013.19
Entropia	-658639.19
MDL	-666389.49

	Accuracy	Precision	Recall	F-Measure	FP-Ratio
Cancellato	79.80%	0.865	0.612	0.717	0.068
Non Cancellato		0.770	0.932	0.843	0.388
Media		0.810	0.798	0.790	0.255

Classificato come ↱	**Cancellato**	**Non Cancellato**
Cancellato	20242	12860
Non Cancellato	3157	43071

Conclusioni e prospettive future

La realizzazione del caso di studio ci ha permesso di comprendere le potenzialità dell'applicazione di metodi di apprendimento automatico in un campo pratico come quello del turismo. Infatti, la costruzione di modelli che permettono la predizione di eventi come la cancellazione di una prenotazione, e l'eventuale numero di giorni che la precedono, possono fornire approfondimenti significativi per i gestori di strutture in tale ambito, ma anche in un qualsiasi contesto inerente alla prenotazione di un servizio. Dal punto di vista progettuale, sarebbe interessante poter testare i medesimi metodi su un più ampio raggio di strutture, magari non limitate alla classe degli hotel "4 stelle" , o con un diverso set di features.

Il caso di studio ha utilizzato un dataset anonimizzato, tuttavia, se fosse stato possibile accedere ad informazioni più caratterizzanti rispetto al tipo di cliente o al contesto dell'albergo in questione, si sarebbe potuto utilizzare fonti di conoscenza esterna come *ontologie*. Oppure, In relazione ai modelli adottati, si sarebbe potuta tentare l'applicazione con delle **Catene di Markov** costruite sui comportamenti dei clienti nel ciclo di vita di una prenotazione o nello storico delle prenotazioni. Un altro spunto di miglioramento potrebbe essere la selezione di un modello alternativo e, magari più adatto, per la predizione sul numero di giorni di preavviso di cancellazione, dati i risultati con ampio margine di miglioramento. In conclusione, è stato osservato che il dominio turistico è un ambito che si presta molto bene ad operazioni relative alla scoperta della conoscenza, dato il grande quantitativo di dati disponibili.