Dipartimento di Impresa e Management

Cattedra di Metodi Statistici per il Marketing

Il Problema della Classificazione Statistica: metodologia ad albero con applicazione al caso PSA Retail Italia

RELATORE

Prof. Pierpaolo D'Urso

CANDIDATO

Micaela Brindisi

Matr. 675851

CORRELATORE

Prof. Carmela Cappelli

ANNO ACCADEMICO 2016/2017

Sommario

INTRODUZIONE

Questo testo affronta il tema della Classificazione Statistica ovvero il problema della predizione di variabili categoriche con l'utilizzo di altre grandezze note (numeriche o categoriche).
I CART, cioè Alberi di Regressione e Classificazione, sono una delle metodologie utilizzate per implementare il tema in esame. Teorizzati empiricamente negli anni '50 principalmente in realtà militare e medica, sono stati successivamente descritti per la prima volta nel 1984 da Breiman tramite la stesura e pubblicazione del libro CART (Classification and Regression Trees), considerato ancora oggi in letteratura, un punto di riferimento.
Vista la crescente complessità dei fenomeni e dell'analisi di dati non trattabili adeguatamente con tecniche tradizionali e di dati non standard (incompleti, per i quali il numero di componenti non risulti fissato o risulti molto elevato), le metodologie ad albero riscuotono sempre più interesse in ambito scientifico grazie anche ai progressi informatici che hanno reso automatica la loro applicazione.

Questo testo è stato strutturato e pensato in 4 Capitoli così suddivisi:

- Capitolo 1: In questo capitolo sono esposte alcune nozioni generali trattanti una serie di temi, propedeutici allo studio dei CART (Classification and Regression Trees).
  È strutturato in una prima parte che introduce l'argomento della Classificazione e le relative caratteristiche, successivamente l'attenzione si concentra sul confronto con la Cluster Analysis descrivendone i principi fondamentali ed in dettaglio la tecnica attualmente più utilizzata, quella del Clustering Gerarchico.

- Capitolo 2: Il secondo Capitolo presenta tutti gli aspetti riguartanti la metodologia ad albero, in una veste teorica presenta il modello, le caratteristiche della metodologia, la costruzione del modello e l'implementazione in R.

- Nel Capitolo 3 è riportata un'analisi effettuata durante la mia esperienza di Stage in PSA Retail Roma grazie a cui ho potuto apprendere come il mondo dell'automotive durante gli ultimi anni sia stato fortemente caratterizzato dall'influenza del Web. Dunque ho analizzato in RStudio un dataset riguardante le richieste online effettuate dagli utenti nel primo trimestre del 2017, utilizzando il pacchetto rpart ( ) per implementre la costruzione dell' albero di classificazione.

- Nella Conclusione sono riportati sinteticamente i giudizi conclusivi relativi all'applicazione del modello.

1. LA CLASSIFICAZIONE

1.1. Classificazione vs Predizione

Quell'ambito della statistica che si serve di informazioni completamente o parzialmente note per dedurre il valore di altre grandezze di nostro interesse invece non note, sfocia nella teoria della predizione.

Parlando di teoria della predizione ci si riferisce prettamente a una nozione generale, rappresentata dalla Classificazione quando la classe è un valore nominale; Regressione quando la classe è un valore numerico. Questi processi hanno come obiettivo la creazione di modelli che permettono di studiare e descrivere gli insiemi di dati e attuare previsioni per il futuro.[1]

Un CART non è nient'altro che un predittore del valore di una variabile di risposta (target) in funzione di un insieme di variabili indipendenti (input).

Utilizzando una terminologia più precisa chiameremo:

• variabili di misurazione le grandezze predittive (**gli input**) che possono essere sia quantitative che qualitative;

• variabile di risposta la grandezza predetta (**l'output**).

Il modello è strutturato secondo un diagramma ad albero e seconda della natura della variabile target, possiamo parlare di:

• Alberi di Regressione (Regression Trees, la variabile target è quantitativa)

• Alberi di Classificazione (Classification Trees, la variabile target è qualitativa).

Parleremo di regressione nel caso in cui la variabile risposta assuma valori continui, di classificazione nel caso in cui i valori assumibili dalla variabile risposta siano categorici (si tratta di classi di appartenenza).

[1] Hastie, T. Tibshirani, R. and Friedman, J. (2001) The elements of statistical learning: data mining, inference and prediction. Springer, New York

1.2. Classificazione e Cluster Analysis

Il termine Classificazione può indicare due problematiche piuttosto simili che storicamente hanno assunto la stessa denominazione ma allo stesso tempo vanno specificate distintamente.

1. Il primo significato del termine Classificazione fa riferimento a quell'insieme di metodologie statistiche che aspirano ad assegnare una classe ad un dato di classe sconosciuta sulla base delle informazioni fornite da un campione di dati di classe invece nota (Learning sample).
2. Il secondo significato del termine Classificazione invece sta ad indicare tutte quelle tecniche che hanno come obiettivo l'individuazione di classi o gruppi più o meno evidenti in un insieme di dati non classificati a priori (Cluster Analysis).

Si nota come le due definizioni siano apparentemente simili ma si differenzino in base ai loro scopi infatti:

- la Classificazione (nel senso di assegnazione) cerca di capire in uno Spazio *X* a quale classe di assegnazione appartenga ogni dato da classificare.
- la Cluster Analysis invece cerca in uno Spazio *X* di capire se esistono delle classi di assegnazione.

Prima di approfondire tutte le problematiche inerenti gli alberi di classificazione è necessario presentare brevemente quali siano le tecniche attualmente più utilizzate.
I metodi statistici per la classificazione delle unità statistiche in gruppi omogenei possono distinguersi in:

- **ANALISI DISCRIMINANTE**
- **CLUSTER ANALYSIS**

Nell'**ANALISI DISCRIMINANTE** (classificazione con supervisione)
Viene utilizzata per spiegare o predire i valori su una singola variabile dipendente categorica utilizzando variabili metriche indipendenti multiple. La variabile target che determina la posizione dei gruppi è di tipo qualitativo (nominale o ordinale).
L'analisi Discriminante è utile per comprendere:

- Le differenze tra gruppi di x variabili;
- Valutare la rilevanza delle x variabili per classificare la variabile y.

L'analisi Discriminante è specialmente utile per capire le differenze e i fattori che portano i consumatoria fare determinate scelte

L'obiettivo dell'Analisi Discriminante è quello di stabilire un criterio che assegni correttamente ulteriori unità alla rispettiva popolazione di appartenenza, minimizzando la probabilità degli errori di attribuzione.

Grazie a questo tipo di analisi è possibile sviluppare strategie di Marketing prendendo in esame il ruolo dei predittori, ad esempio:

- Direct Mail response model (risposta / non risposta);
- Segmentare e profilare acqurenti;
- Determinare la Customer Loyalty (molto leale, mediamente leale, non leale);
- Individuare i fattori che portano all'acquisto o al non acquisto.

La **CLUSTER ANALYSIS** (classificazione senza supervisione) è un metodo esplorativo che consiste nel ricercare nelle n osservazioni p-dimensionali gruppi di unità tra loro simili, non essendo nota a priori l'esistenza nel dataset di tali gruppi omogenei e le caratteristiche strutturali di eventuali gruppi.

La Cluster Analysis ha come obiettivo riconoscere gruppi che appaiono con "naturalezza" nelle osservazioni[2].

1.3. CLUSTER ANALYSIS

1.3.1. Aspetti generali e approcci metodologici

Una prima descrizione sistematica della Cluster Analysis è opera di Tryon (1939) e negli ultimi decenni, grazie allo sviluppo del settore informatico, sono stati risolti molti problemi computazionali connessi a questi metodi.

[2] Metodi Statistici per il Marketing, Prof. Pierpaolo D'Urso, 2016

Questi sviluppi hanno condotto a nuovi algoritmi operativi, facilmente implementabili su PC e di conseguenza tali procedure sono state utilizzate in vari ambiti applicativi (economia, marketing, sociologia, psicologia, medicina, antropologia, ecc.).

La CLUSTER ANALYSIS è una tecnica statistica multivariata che ha come obiettivo l'individuazione di una o più partizioni in gruppi detti CLUSTER dell'insieme di n unità statistiche.

I CLUSTER (a due a due disgiunti in base ad un set di variabili) devono essere caratterizzati da:

1. **COESIONE INTERNA** (le unità assegnate ad uno stesso gruppo devono essere omogenee all'interno di ciascun gruppo).
2. **SEPARAZIONE ESTERNA** (i gruppi devono essere il più possibile eterogenei tra loro).

Questo tipo di Analisi viene usata principalmente per:

- ridurre i dati (nel senso delle unità),
- identificare tipologie,
- stratificare popolazioni da sottoporre a campionamento,
- segmentare il mercato e studiare il comportamento della clientela,
- individuare aree omogenee (geomarketing e aree-test di mercato).

Contrariamente all'Analisi Discriminante, non si sa nulla a priori sulle caratteristiche strutturali dei gruppi.

Nei metodi di clustering ogni unità statistica appartiene ad uno ed un solo cluster, diversamente dai metodi di clustering non tradizionali come i metodi di clustering con overlapping (clumping) e i metodi di clustering fuzzy, in cui ogni unità può appartenere a più cluster.

1.3.2. Clustering Gerarchico

I metodi di Clustering si distinguono in **METODI GERARCHICI e METODI NON GERARCHICI,** in questo elaborato prenderemo principalmente in esame i primi in quanto permettono di ottenere una famiglia di partizioni.

Possiamo suddividere i **metodi gerarchici** in:

- **metodi gerarchici aggregativi** con un numero di gruppi da n a 1 in cui tutte le unità sono distinte per giungere, per aggregazioni successive, a quella banale in cui tutte le unità sono riunite in un unico gruppo,[3]
- **metodi gerarchici scissori** con un numero di gruppi da 1 a n, partendo da quella banale in cui tutte le unità sono riunite in un unico gruppo per giungere, per separazioni successive, a quella anch'essa banale, in cui tutte le unità sono distinte in n gruppi.

I **metodi non gerarchici** forniscono un'unica partizione delle n unità in g gruppi, con g fissato a priori.

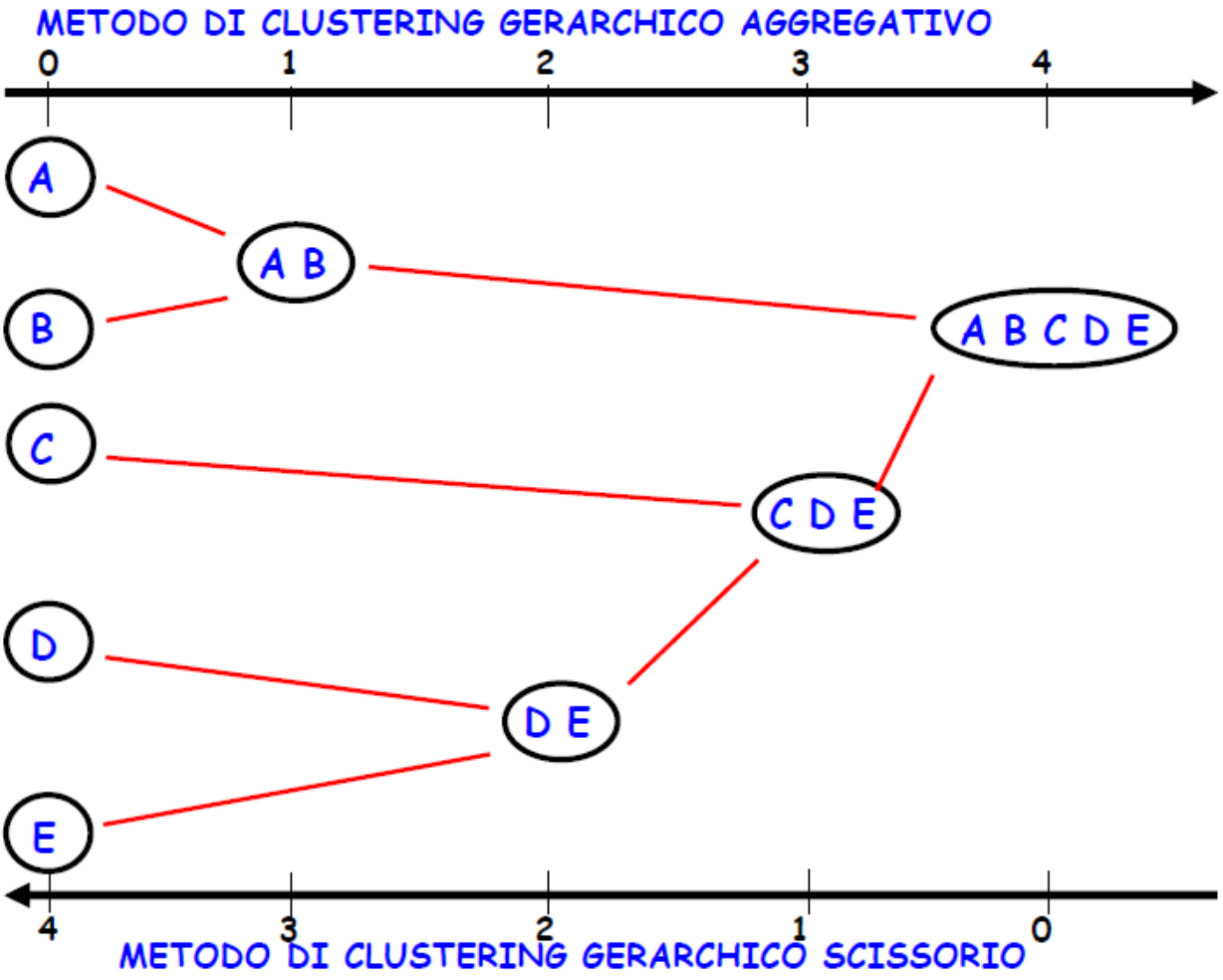

Figura 1-Metodi di Clustering Gerarchico- Metodi statistici per il Marketing, prof. Pierpaolo D'Urso, 2016

[3] Metodi Statistici per il Marketing, Prof. Pierpaolo D'Urso, 2016

Il metodo di Clustering Gerarchico Aggregativo parte con la trattazione della Matrice dei Dati del tipo *unità* × *variabili*, di dimensione n × p (in forma compatta):

$$\mathbf{X} = \{ x_{is}: i = 1,\ldots, n; s = 1,\ldots,p\}.$$

La Matrice dei Dati può essere trasformata con opportune procedure di centratura, normalizzazione, standardizzazione (data pre-processing).

Possiamo Considerare la Matrice degli scarti standardizzati, al fine di rendere comparabili variabili originariamente espresse in unità di misura differenti e con diverso ordine di grandezza.

$$Z = \{z_{is} = \frac{x_{is} - \bar{x}_s}{\sigma_s} : i = 1,\ldots,n; s = 1,\ldots,p\}$$

[z_{is} rappresenta lo scarto standardizzato per l'unita i-esima e la variabile s-esima,

$\bar{x}_s$=media, σ_s =s.q.m.]

Tuttavia, una vasta classe di metodi gerarchici (aggregativi) si basa sull'utilizzo di una una matrice simmetrica di dimensione n × n chiamata Matrice delle Prossimità calcolata per le n unità statistiche:

$$\mathbf{P} = \{p_{ij}: i, j = 1,\ldots,n; i \neq j\}$$

[*pij* rappresenta una misura di prossimità tra la coppia di unità *i* e *j*].

In particolare possono costruirsi Matrici delle Distanze, Matrici delle Dissimilarità, Matrici delle Similarità.

I metodi di Clustering gerarchici aggregativi comprendono due particolari caratteristiche:

1. Considerano tutti i livelli di distanza γ , $0 \leq \gamma \leq \infty$;

2. I gruppi che si ottengono ad un livello di distanza comprendono i gruppi ottenuti ai livelli inferiori. Quindi quando 2 (o più) unità si uniscono tra loro, non possono essere separate nei passi successivi della procedura.

Le differenze tra i vari metodi gerarchici (aggregativi) consistono nel criterio utilizzato per calcolare la distanza tra 2 gruppi di unità (uno dei quali eventualmente formato da una sola unità).

I più diffusi metodi di Clustering Gerarchici Aggregativi sono:

1. METODO DEL LEGAME SINGOLO (O SINGLE LINKAGE O DEL VICINO PIU' PROSSIMO)
2. METODO DEL LEGAME COMPLETO (O COMPLETE LINKAGE O DEL VICINO PIU' LONTANO)
3. METODO DEL LEGAME MEDIO (TRA I GRUPPI) (O AVERAGE LINKAGE)
4. METODO DEL LEGAME MEDIO NEI GRUPPI
5. METODO DI WARD (O DELLA MINIMA DEVIANZA)
6. METODO DEL CENTROIDE

Se indichiamo con:

- R e S due generici cluster non aventi individui in comune;
- N_R e N_S rispettivamente il numero di individui che costituiscono i due cluster;
- $\mathbf{x_{Ri}}$ e $\mathbf{x_{Sj}}$ rispettivamente le coordinate dell'i-esimo individuo del cluster R e del j-esimo
  individuo del cluster S;
- $d(\mathbf{x_{Ri}}, \mathbf{x_{Sj}})$ la distanza tra i sopra citati individui;

potremo indicare le diverse distanze intercluster nel seguente modo:

• **Single linkage**

$lksingle\ (R,S) = min_{(ij)}(d(\mathbf{x_{Ri}}, \mathbf{x_{Sj}}))$ $\quad i \in \{1,2,\ldots, N_R\}$ e $j \in \{1,2,\ldots, N_S\}$

la distanza tra R e S coincide con la minima distanza che intercorre tra gli individui dei due cluster.

Figura 2- Single Linkage- Metodi statistici per il Marketing, prof. Pierpaolo D'Urso, 2016

• **Complete linkage**

lkcomplete (*R*,*S*) = = $max_{(ij)}(d(\mathbf{x_{Ri}}, \mathbf{x_{Sj}}))$ i∈{1,2,…, N_R }e

j∈{1,2,…, N_S }

la distanza tra *R* e *S* coincide con la massima distanza che intercorre tra gli individui dei due cluster.

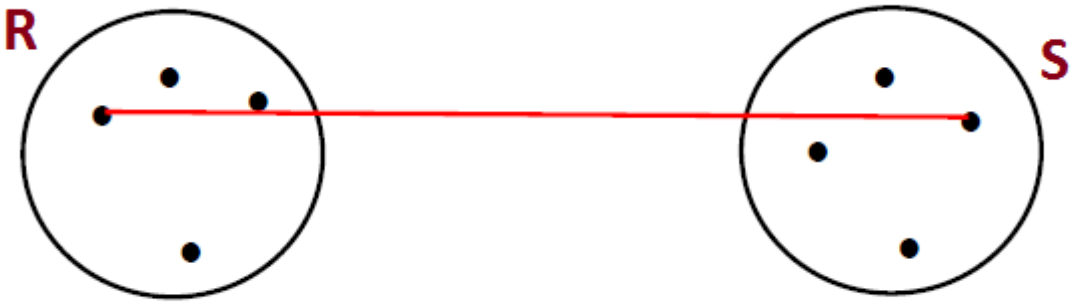

Figura 3- Complete Linkage- Metodi statistici per il Marketing, prof. Pierpaolo D'Urso, 2016

• **Average linkage**

$$lkaverage\ (R,S) = \frac{\sum_{i=1}^{N_R}\sum_{j=1}^{N_S} d(\mathrm{x}_{Ri}\ ,\mathrm{x}_{Sj})}{N_R N_S}$$

la distanza tra *R* e *S* coincide con la distanza media che intercorre tra gli individui dei due cluster.

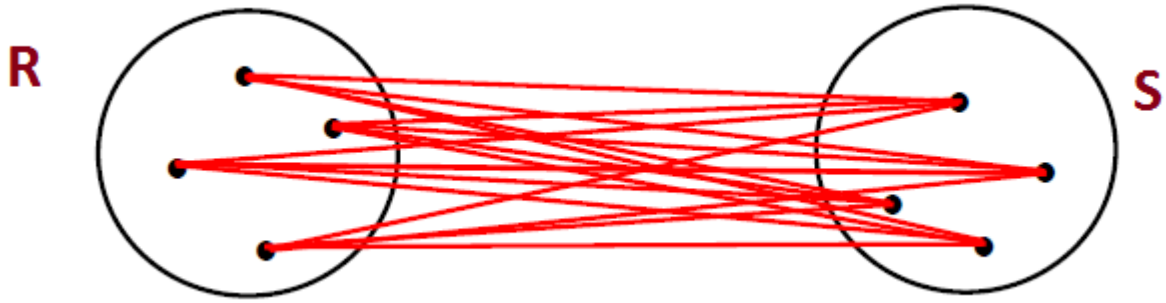

Figura 4- Average Linkage-Metodi statistici per il Marketing, prof. Pierpaolo D'Urso, 2016

• **Centroide linkage**

lkcentroide (*R*,*S*) = $d(\mathbf{x}_R, \mathbf{x}_S)$

$$\text{con } \mathbf{x}_R = \frac{\sum_{i=1}^{N_R} \mathbf{x}_{Ri}}{N_R} \; ; \; \mathbf{x}_S = \frac{\sum_{j=1}^{N_S} \mathbf{x}_{Sj}}{N_S}$$

la distanza tra *R* e *S* coincide con la distanza che intercorre tra i baricentri $\mathbf{x}_R$, e $\mathbf{x}_S$ dei due cluster.

Notiamo che a causa della non linearità della funzione distanza *d*(,) avremo tipicamente che la distanza tra i baricentri *lkcentroide* (*R*,*S*) sarà differente dalla distanza media *lkaverage* (*R*,*S*).

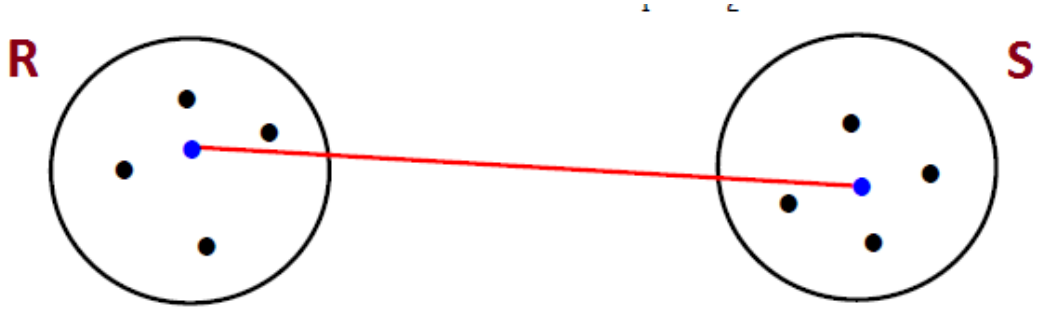

Figura 5- Centroide Linkage- Metodi statistici per il Marketing, prof. Pierpaolo D'Urso, 2016

• **Ward linkage**

$$lk^2_{Ward}\ (\mathrm{R,S}) = \frac{N_R N_S}{N_R + N_S}\, lk^2_{centroide}\ (\mathrm{R,\ S})$$

Per capire da dove nasca l'idea di una tale distanza è necessario introdurre la cosiddetta somma quadratica per gruppi cioè la somma dei quadrati delle distanze di ogni individuo dal baricentro del proprio cluster di appartenenza. Questa grandezza è calcolabile per ogni partizione e può variare tra zero, nel caso della partizione banale P_N costituita da N cluster con un solo individuo, alla varianza statistica della popolazione moltiplicata per (N-1), nel caso di un unico cluster comprendente tutti gli N individui.

In riferimento al clustering gerarchico si può dimostrare che fondendo due cluster *R* e *S* la somma quadratica per gruppi non può che aumentare e che tale incremento coincide proprio col quadrato della distanza di Ward *lkWard* (*R*,*S*) che intercorre tra i due cluster *R* e *S*.

Quindi secondo la distanza di Ward diremo che due cluster sono tanto più vicini quanto minore è l'incremento della somma quadratica per gruppi generata dalla loro eventuale fusione.

In conclusione il metodo del legame singolo e il metodo del legame completo individuano cluster con caratteristiche diverse, infatti con il metodo del legame singolo è possibile classificare in un unico cluster unità molto distanti, nel caso in cui tra esse esista una

successione di punti (quindi di unità) intermedi, verificandosi di conseguenza il cosiddetto EFFETTO CATENA.
Il metodo del legame completo individua invece CLUSTER COMPATTI DI FORMA CIRCOLARE (in R^2) (o SFERICI in R^3 o IPERSFERICI in R^p).

Figura 6- Effetto Catena- Metodi statistici per il Marketing, prof. Pierpaolo D'Urso, 2016

Nella Figura 5 è riportato un esempio di come l'effetto catena determini la presenza di cluster diversi a seconda del metodo utilizzato.
Chiaramente nota è la presenza di 3 Cluster tuttavia l'esistenza di alcuni punti tra i due cluster a sinistra fa in modo che (per l'effetto catena) il metodo del legame singolo individui 2 cluster (unificando i 2 cluster a sinistra), mentre il metodo del legame completo ne individui, in modo corretto, 3 (assegnando equamente i punti intermedi tra i due cluster) nonostante l'effetto catena del legame singolo abbia il vantaggio di individuare cluster con forme allungate.
In conclusione, il metodo del legame medio può costituire un giusto compromesso tra metodo del legame singolo e del legame completo.

1.3.3. Rappresentazione grafica dei risultati ottenuti con metodi di Clustering Gerarchici Aggregativi

La famiglia di partizione ottenuta con un metodo gerarchico aggregativo può essere rappresentata graficamente mediante un ALBERO n-DIMENSIONALE (n-TREE).

In particolare, si considera un sistema di assi cartesiani in cui in ascisse si pongono le unita statistiche e in ordinata gli STADI della classificazione.

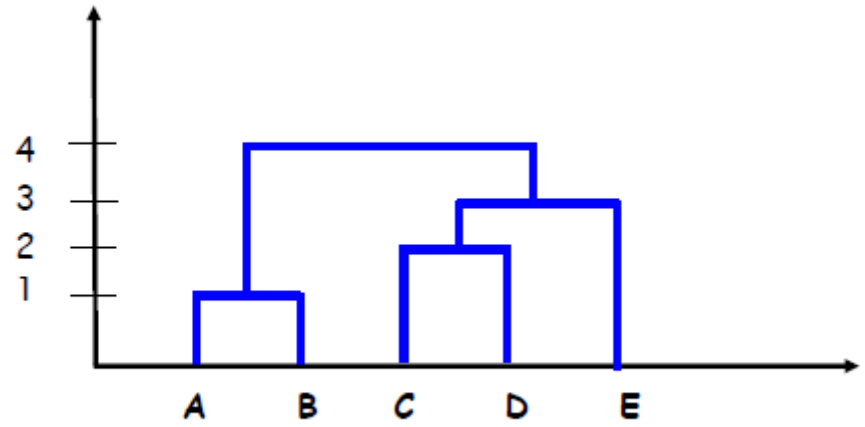

Figura 7- Esempio di albero n-dimensionle- Metodi statistici per il Marketing, prof. Pierpaolo D'Urso, 2016

Al fine di ottenere un'analisi più precisa possiamo considerare in ordinata i livelli di distanza che caratterizzano le aggregazioni delle diverse partizioni. I tal modo si ottiene una successione di partizioni annidate fondendo in un unico nuovo cluster due cluster della partizione precedente. La successione così ottenuta può essere rappresentata graficamente tramite un DENDOGRAMMA cioè una struttura ad albero nella quale ogni ramo rappresenta un possibile cluster ed ogni potatura effettuata "parallelamente al suolo" una partizione della successione (una struttura che graficamente risulterà simile a quella di un albero di classificazione ma concettualmente molto diversa).

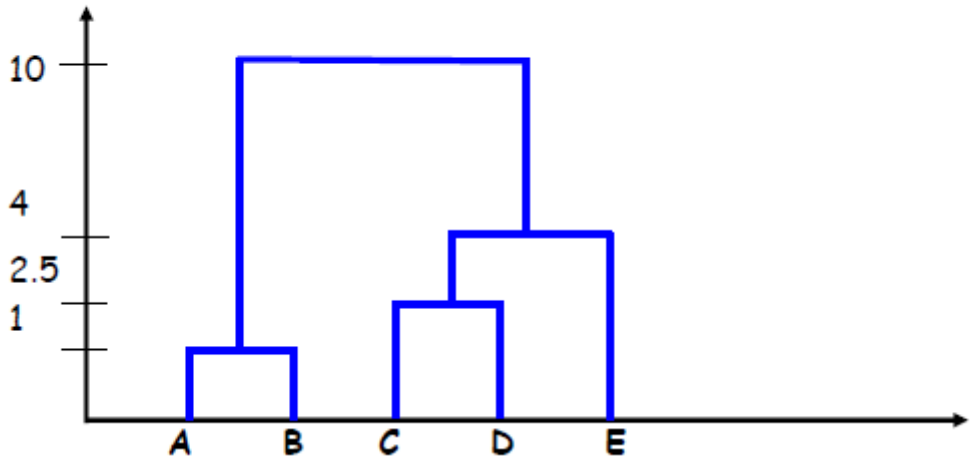

Figura 8- Dendogramma- Metodi statistici per il Marketing, prof. Pierpaolo D'Urso, 2016

Un algoritmo gerarchico aggregativo genera una famiglia di partizioni delle n unità con un numero di cluster via via decrescenti da n a 1, si pone quindi in essere il problema di scegliere, tra le diverse partizioni, la partizione ottimale (e quindi il numero di cluster ottimale).

Considerando le condizioni desiderabili di coesione interna (omogeneità all'interno dei cluster) e separazione esterna (eterogeneità tra i cluster) e quindi la scomposizione della devianza totale, si possono distinguere diversi criteri per determinare il numero ottimo di Cluster:

- TEST DI SEPARAZIONE TRA CLUSTER: si considerano test per verificare se la distanza tra i centroidi dei cluster sia significativa.
- INDICI SINTETICI**:** ad esempio indice R^2, indice RMSSTD, indice C di Calinski Harabatz, ecc. Una "buona" classificazione è caratterizzata da una ridotta quota di devianza nei cluster (W) e da un elevato valore delle devianza tra i cluster (B). Quindi, data una partizione costituita da g cluster, si può considerare l'indice:

  $R^2 = 1 - \frac{W}{T} = \frac{B}{T}$ con $R^2 \in [0,1]$.

  Se R^2è prossimo a 1 la corrispondente classificazione può ritenersi omogenea poichè le unità appartenenti ad un medesimo cluster sono simili tra loro (W≅ 0) e i cluster sono ben separati (B ≅T). Esiste un trade-off tra il numero di cluster e la coesione interna, in particolare, R^2 assume valori non decrescenti al crescere

di g (numero di cluster). Quindi, la ricerca del numero "ottimo" di cluster non può fondarsi semplicemente sulla massimizzazione di R^2 ma deve compendiare le esigenze contrapposte di omogeneità interna dei cluster e di sintesi della classificazione.

- ISPEZIONE (DIRETTA) DEL DENDOGRAMMA (α-TAGLIO): si "taglia" il dendogramma in corrispondenza di un "salto" nei livelli di distanza in cui è avvenuta l'aggregazione.
  Ad esempio

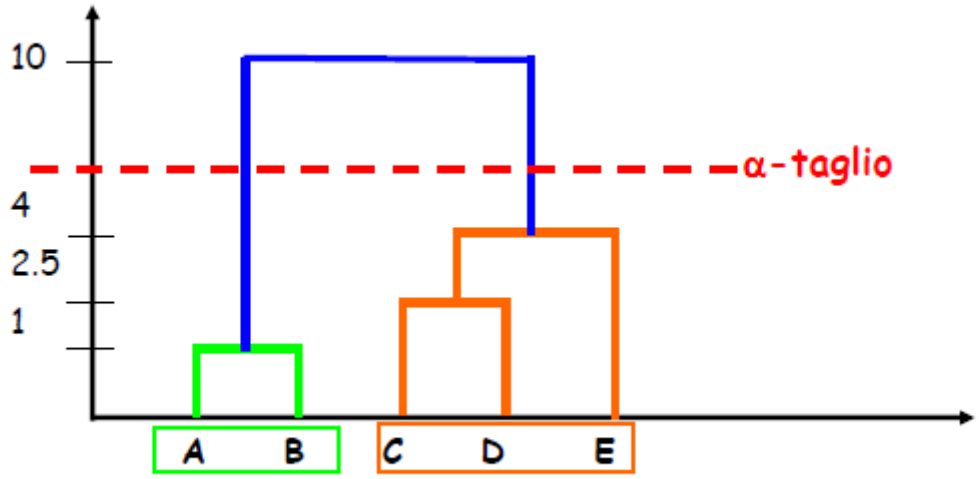

Figura 9- α-TAGLIO- Metodi statistici per il Marketing, prof. Pierpaolo D'Urso, 2016

Si calcolano i vettori medi di ciascun cluster ottenuto e si confrontano i valori dei vettori riguardanti le variabili in esame, dunque la caratterizzazione di ogni cluster avviene in base a tali valori.

1.3.4. Altri Metodi di Clustering Gerarchici Aggregativi

Possiamo classificare due tipologie alternative di Clustering Gerarchico Aggregativo, distinguiamo:

• Metodi di clustering vincolati

• Metodi di clustering con parziali sovrapposizione.

Con il termine **Classificazione vincolata** si indicano i metodi per ottenere gruppi di unità che risultino simili tra loro relativamente alle variabili in esame, e che soddisfino anche ulteriori condizioni.
Possiamo definire diverse tipologie di vincoli; in particolare:

- Vincoli sul numero di unità assegnate ad ogni cluster; formato da un numero di unità comprese in un certo intervallo, i cui estremi siano prefissati dal ricercatore;
- Vincoli di contiguità (nel caso in cui le unità da classificare siano di tipo territoriale);

Un tipo di vincolo molto utilizzato nell'analisi di dati territoriali o spaziali è quello di CONTIGUITA' che viene ben definito, tra le unità assegnate ad ogni cluster. Tale vincolo è essenziale in molti problemi di analisi territoriale, in cui le unità (comuni, province, regioni, mercati locali del lavoro, etc.) di un gruppo devono essere non solo simili, ma anche tra loro vicine.
[Il concetto di contiguità è riferito al tipo di unità territoriale e alla dimensione dello spazio considerato.]

Considerato uno spazio bidimensionale R^2 e che le unità territoriali siano collocate su una superficie, possiamo distinguere il caso di **unità territoriali di forma REGOLARE e IRREGOLARE.**
Due maglie quadrate appartenenti ad un reticolo regolare risultano avere un lato o un vertice in comune, sono considerate contigue tra loro e si parlerà di unità territoriali i form regolare, viceversa si assume abitualmente che siano contigue due unità territoriali con una porzione di confine in comune.
Dopo aver definito la contiguità tra 2 unità territoriali, si costruiscono i diversi cluster (aree omogenee) applicando un metodo di clustering (ad esempio, gerarchico) vincolato:

- due unità territoriali possono essere assegnate allo stesso cluster solo se sono tra loro spazialmente contigue;
- un'unità territoriale può essere assegnata ad un cluster precedentemente formato se essa è contigua con almeno una delle unità del cluster stesso;
- due unità si possono unire tra loro se almeno una unità di un cluster è contigua ad una unità dell'altro cluster.

Formalmente, il vincolo di contiguità (territoriale) viene considerato nelle procedure di clustering affiancando alla matrice delle distanze D una matrice di contiguità (detta anche matrice di adiacenza), simmetrica, di dimensione n×n e con valori unitari lungo la diagonale principale, così definita:

$$\mathbf{C}=[c_{ij}]$$

dove

$$c_{ij} = \begin{cases} 1 \\ 0 \end{cases}$$

L'introduzione del vincolo di contiguità nei metodi di clustering (gerarchici) può produrre l'inconveniente dell' INVERSIONE nei valori della distanza in corrispondenza della quale una unità territoriale si unisce ad un cluster e si dimostra che il metodo del legame completo con il vincolo di contiguità non produce inversione.

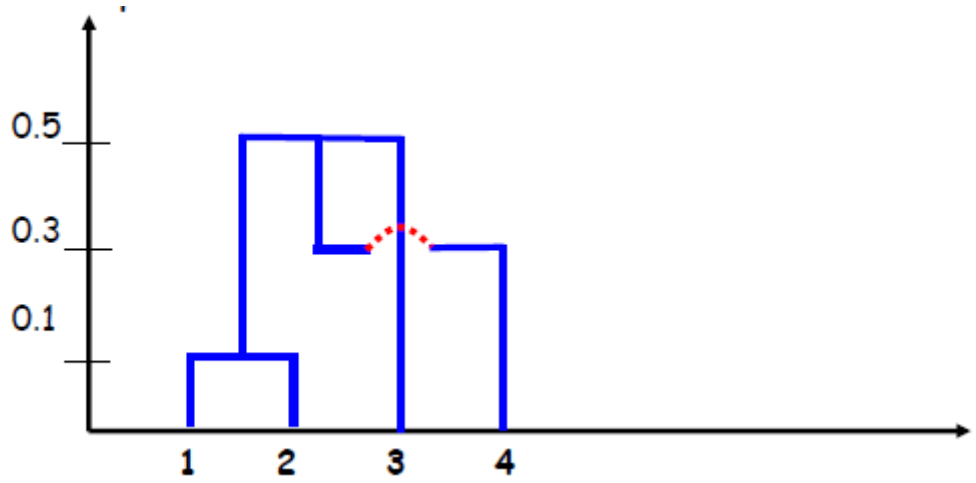

Figura 10- Inversione dei valori nella distanza- Metodi statistici per il Marketing, prof. Pierpaolo D'Urso, 2016

Differentemente, i metodi di Clumping sono metodi di Clustering che generano classificazioni con parziali sovrapposizioni.

Eccessive sovrapposizioni possono portare a un'interpretazione dei risultati oscura e può rendere inutile la classificazione. Occorre, quindi, introdurre opportuni vincoli per limitare le sovrapposizioni consentite. Ad esempio, nel cosiddetto **metodo Bk** (k=1,2,3, …) il numero massimo di unità che possono sovrapporsi in ciascuna coppi di gruppi e

uguale a (k-1); se il numero di unità in comune è maggiore, i due gruppi vengono fusi in uno solo. (Per k=1 tale metodo coincide col legame singolo).

I metodi di clumping sono molto applicati nella **classificazione di aree territoriali** (**zonizzazione**=**zoning**): le unità territoriali che appartengono contemporaneamente a due gruppi rappresentano gli elementi "**di cerniera**"[4] tra le aree omogenee individuate, con caratteristiche intermedie a quelle dei gruppi corrispondenti.

Una **rappresentazione grafica** della gerarchia di partizioni ottenute attraverso l'utilizzo di un metodo di clumping di tipo gerarchico è costituita dal cosiddetto **DENDROGRAMMA A PIRAMIDE**.

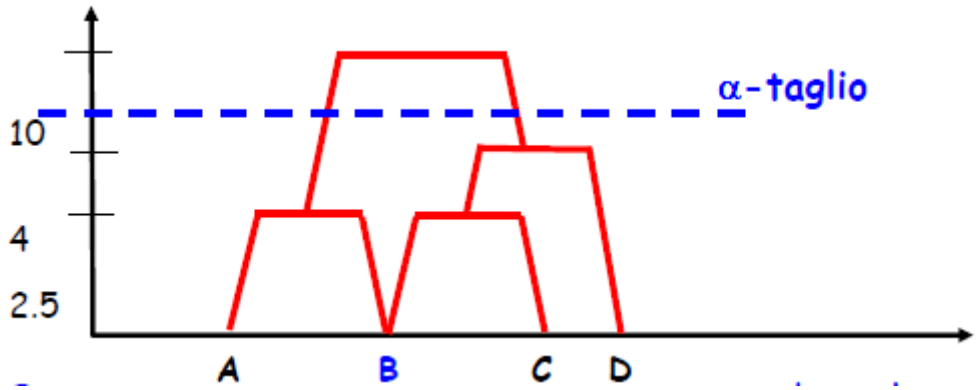

Figura 11- Dendogramma a Piramide- Metodi statistici per il Marketing, prof. Pierpaolo D'Urso, 2016

In tal caso l'unita **B** appartiene contemporaneamente a due cluster: **[A,B]** e **[B,C,D]**.

2. GLI ALBERI DI CLASSIFICAZIONE

2.1. Introduzione al modello

I modelli strutturati ad albero sono stati usati in campi diversi come la botanica e la medicina, prima che il loro potenziale esplicativo venisse scoperto dagli statistici, specialmente per applicazioni relative a problemi di classificazione.

Gli alberi di classificazione (o di s*egmentazione*) rappresentano una metodologia che ha l'obiettivo di ottenere una segmentazione gerarchica di un insieme di unità statistiche

[4] Metodi statistici per il Marketing, prof. Pierpaolo D'Urso, 2016

mediante l'individuazione di "regole" (o "percorsi") che sfruttano la relazione esistente tra una classe di appartenenza e le variabili rilevate per ciascuna unità, dunque sono il risultato di un processo di suddivisione ricorsiva di un insieme di unità statistiche in sottogruppi disgiunti, caratterizzati da un grado di omogeneità crescente e di numerosità via via inferiore[5].

Formalmente un albero è costituito da un insieme finito di elementi detti **nodi**; il nodo da cui si diramo i successivi viene detto **radice**.

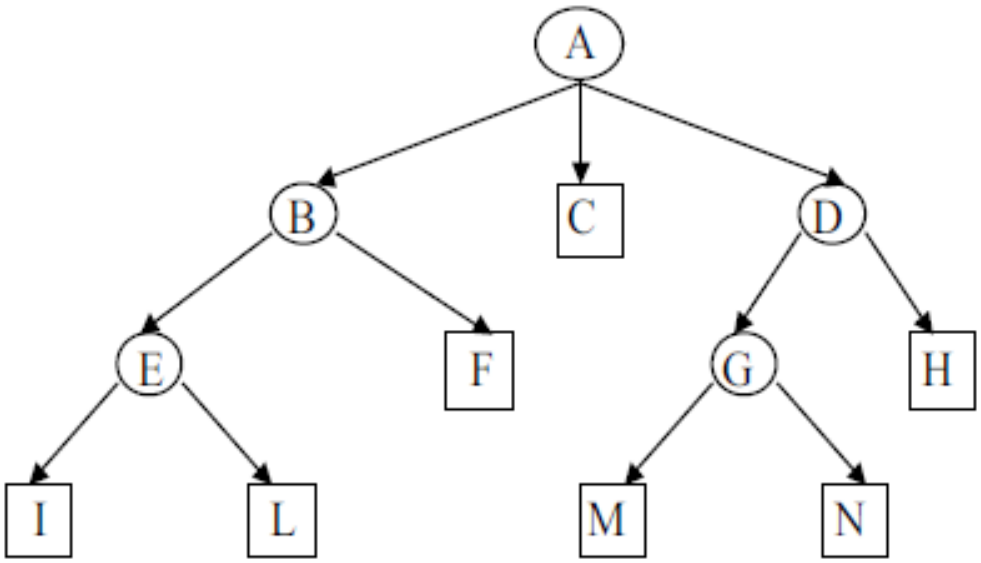

Figura 12- Esempio di generica struttura ad albero- Non parametric regression and classification theroy and an application to web-marketing- Prof. C.Cappelli, 2016

L'albero rappresentato in *Figura 12* presenta 4 livelli, partendo dal nodo radice (A) che costituisce il primo livello, si incontrano due nodi *interni* (B,D) nel secondo livello insieme ad un nodo cosiddetto *terminale* o *foglia* indicato dal riquadro (D). Passando al terzo livello sono riportati ancora una volta due nodi interni (E,G) e due terminali (F,H)

[5] Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J.(1984). Classification and regression trees, Wadsworth International

ed infine nel quarto livello sono rappresentati i nodi terminali (I, L, M, N) con cui termina la struttura non essendo ulteriormente suddivisi. Si noti come ciascun nodo interno, rappresentato da un cerchietto, sia a sua volta radice di un sottoalbero o *branca*: potando la branca che deriva da un qualunque nodo interno, si ottiene un sottoalbero dell'albero originario, che presenta la stessa radice.

Gli split sono i valori soglia che dividono le unità di un determinato nodo. Il modo più semplice ed efficace per classificare osservazioni in un numero finito di classi sono gli alberi di decisione che vengono realizzati suddividendo più volte le osservazioni in sottoinsiemi omogenei rispetto alla variabile dipendente o risposta.

Questo metodo crea una gerarchia ad albero, dove i sottoinsiemi iniziali sono i nodi(e quelli finali le foglie.

Nello specifico, i nodi sono etichettati con il nome delle variabili, gli archi, cioè i rami dell'albero, con i possibili valori della variabile soprastante, e le foglie dell'albero con le diverse modalità della variabile "classe" che descrivono i gruppi di appartenenza.

Così facendo ogni oggetto viene classificato seguendo un determinato percorso lungo l'albero, dalla radice ad una precisa foglia. I percorsi sono rappresentati dai rami dell'albero che forniscono le regole base per orientarsi nel sistema.

Dalla Figura 11 notiamo come ciascun nodo interno dia luogo ad un numero variabile di suddivisioni; nel caso il numero delle variabili sia pari a 2 e resta costante per ogni nodo interno e per ogni livello, ci troviamo difronte a degli alberi binari, importanti perché le metodologie ad albero tipicamente operano per dicotomie all'interno del dominio fondamentale in esame dando luogo ad alberi di tal tipo, pertanto dette di segmentazione binaria. [6]

Dall'analisi appena fatta è evidente come il ruolo degli alberi di decisione sia importante tanto quanto utile per il supporto di determinate decisioni e la rappresentazione di diversi tipi di informazione; lampante è l'esempio dell'albero genealogico anche se moltissime altre informazioni sono rappresentate con una struttura ad albero e tante altre si presterebbero alla medesima rappresentazione Questo tipo di strumento ha il vantaggio di esprimere graficamente i concetti di progressività e di inclusione e di conseguenza di gerarchizzazione, infatti consente di rappresentare legami gerarchici tra i dati.

[6] Buntine, W. (1992). Learning classification trees, Statistics and Computing, 2, 63-73.

2.2. Le metodologie ad albero: notazione e concetti preliminari

Le metodologie ad albero sono utili ai fini di predire la variabile *Y* detta *variabile di risposta* o *criterio* sulla base dei valori assunti da un insieme di predittori *X*1*;X*2*; : : : ;XK*.

Nel caso in cui la variabile di risposta sia *categorica*, riferendoci alle metodologie ad albero possiamo parlare di metodo di Classificazione, diversamente in caso di variabile criterio di tipo *numerico* si parla di *Regressione ad albero*.

In entrambi i casi il fine ultimo non è altro che l'assegnazione alla variabile *Y* di un valore che rappresenta un'etichetta corrispondente ad una classe.

Tecnicamente, si prende in esame la probabilità:

$$f\,(Y = y|\, X_1, X_2, \ \dots\ , X_K); \qquad (2.2.1)$$

Nel caso in cui *Y* sia numerica si considera il valore atteso:

$$E(Y \mid X_1, X_2, \ \dots\ , X_K). \qquad (2.2.2)$$

In ambito statistico se esaminassimo la variabile *Y* dicotomica o politomica, ricorreremmo alla regressione logistica [7], e nel caso in cui i predittori fossero tutti di tipo numerico utilizzeremmo l'analisi Discriminante lineare di Fisher o le sue estensioni e varianti quali ad esempio l'analisi discriminante quadratica, che considera interazioni di ordine superiore tra i predittori, infine se prendessimo in esame una variabile di risposta numerica la soluzione standard verrebbe fornita dal modello di regressione lineare.

Tali metodi tradizionali si fondano su ipotesi fortemente restrittive riguardanti la forma distribuzionale dei dati o il tipo di legame esistente tra la variabile di risposta ed i predittori. Di conseguenza si otterrebbero risultati poco affidabili con un'interpretazione dei dati e del modello in esame decisamente problematica.

In contrasto con tali metodi classici, le metodologie ad albero presentano dei notevoli vantaggi:

- Essendo tecniche non parametriche non necessitano di un modello specifico;

[7] In tal caso la (2.2.2) coincide con la (2.2.1) con y= 1.

- I predittori da utilizzare possono risultare di diversa natura;
- La rappresentazione grafica che ne scaturisce risulta di facile interpretazione in quanto consente di visualizzare le relazioni esistenti tra variabile di risposta e predittori in maniera immediata.

Si potrebbe dire che tali metodologie rispondono ad un problema classico della statistica senza presentare molti degli inconvenienti dei metodi classici impiegati al medesimo scopo. [8]

2.3. Gli alberi di classificazione: caratteristiche della metodologia

Come anticipato nel paragrafo precedente il sistema di rappresentazione ad albero di classificazione, può avere svariati utilizzi perché la loro struttura ha il vantaggio di poter essere memorizzata in modo uniforme, di aggiungere alla classificazione eventuali nuove informazioni, e sono applicabili per risolvere problemi di diversa natura.

Tra gli svantaggi invece troviamo la difficoltà nel trovare un albero di dimensione ottimale nonostante esistano metodi "top-down" che partono dalla radice ripartendo solo successivamente lo spazio delle variabili.

Per costruire una struttura ad albero efficace occorre seguire 3 step:

1. Selezionare una regola di splitting per ogni nodo, ciò significa determinare le variabili, insieme al rispettivo valore soglia, che saranno usate per partizionare il data set ad ogni nodo.
2. Determinare quali nodi sono da intendersi terminali, quindi per ogni nodo bisogna decidere quando continuare con gli splits, quando fermarsi e considerare il nodo come terminale e di conseguenza assegnargli un'etichetta. Infatti senza un'adeguata regola, si corre il rischio di costruire alberi troppo grandi con una piccola capacità di generalizzazione, oppure alberi troppo piccoli che invece approssimano male i dati.

[8] Cappelli C. (2002), La Validazione della struttura nelle metodologie ad albero. Tesi di Dottorato di Ricerca, XII ciclo. Università di Napoli Federico II

3. Assegnare le etichette ad ogni nodo terminale, ad esempio minimizzando il valore atteso di errata classificazione. A partire da un qualunque tipo di problema è quasi sempre possibile costruire l'albero di decisione corrispondente.

Lo scopo finale è l'individuazione di relazioni esistenti tra la variabile di risposta ed i predittori, cosicché sia necessario individuare un insieme di regole di classificazione/predizione, nella forma di un albero binario.Viene utilizzato un insieme di apprendimento, C = {(y_i, x_i, i = 1, … , n)} ovvero un insieme di osservazioni su cui sono state rilevate i valori assunti dalla variabile criterio e dai predittori ; l'insieme di casi si supponga essere estratto da una variabile casuale multivariata (X; Y) dove X è il vettore dei K predittori ed Y è la variabile di risposta.

Questo insieme rappresenta l'esperienza passata e costituisce ciò che si conosce, il cosiddetto dominio fondamentale, impiegato appunto per creare l'albero, con la duplice valenza, di esplorare i dati dell'insieme di apprendimento e di indurre il valore di risposta da associare a nuove osservazioni che scivolando nella struttura raggiungono un nodo terminale.

Si consideri un esempio di output di dati raccolti negli USA, in cui la variabile di risposta (y) è il reddito annuale, suddiviso in due classi: >50.000$ e <50.000$.

Le variabili esplicative sono: l'età (anni passati a scuola), relazione familiare, e il capital gain.

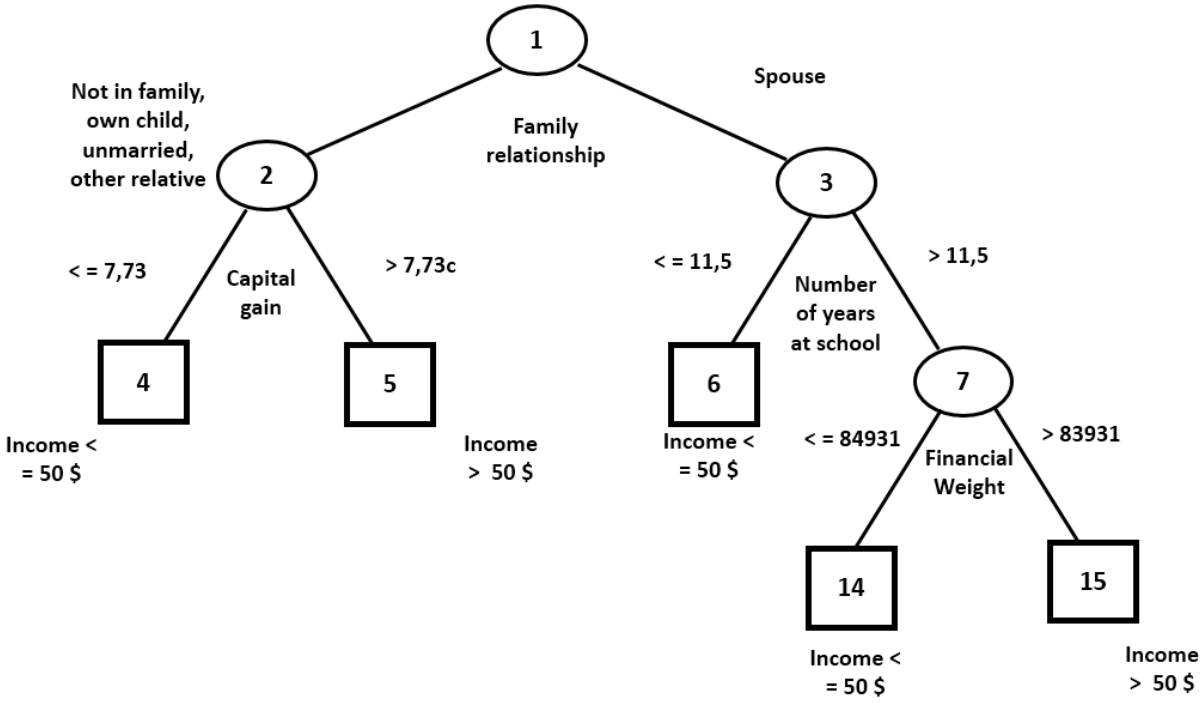

Figura 13- Albero binario Ipotetico- Non parametric regression and classification theroy and an application to web-marketing- Prof. C.Cappelli, 2016

All'inizio della procedura nel nodo 1 che è il nodo radice da cui tutta la struttura nasce, sono presenti tutti i rispondenti, dunque tutte le osservazioni.

Si nota come la variabile (nell'ambito delle variabili esplicative) che maggiormente suddivide queste osservazioni in due sottogruppi più o meno omogenei rispetto al reddito annuale è quella delle Relazioni Familiari. Nel **nodo 2** (nodo figlio a sinistra) vanno a finire tutti i rispondenti che non sono in famiglia, *unmarried*, o vivono con altri familiari; nel **nodo 3**, sono presenti i rispondenti in una condizione *spouse*, coniugale. Quindi in base alla variabile *relazione familare* i rispondenti sono stati divisi in due sottogruppi che sono più omogenei rispetto all' unico insieme iniziale e sono omogenei in termini di reddito annuale che percepiscono.

Nel **nodo 3** la variabile che risulta maggiormente esplicativa, che suddivide i rispondenti *spouse* in sottogruppi ancora più omogenei, è il numero di anni a scuola.

In particolre, nel **nodo 6** abbiamo tutti i rispondenti in condizione *spouse* che però non hanno un grado di istruzione elevata, e queste persone percepiscono un reddito modale < 50.000$. Quelli che invece hanno raggiunto più della soglia di 11,5 in termini di

numero di anni scolastici finiscono nel **nodo 7** in cui la variabile che li discrimina in sottogruppi ancora più omogenei è *financial weight*

Questa struttura grafica quindi sulla base delle covariate ci dà una serie di possibili regole di predizione, e queste regole sono 5, quanti sono i nodi terminali. Perché corrispondono a diversi percorsi che dal nodo radice conducono a un gruppo omogenei rispetto alla variabile di risposta.

Si noti come i nodi siano designati secondo una numerazione binomia che, a partire dal nodo radice cui è assegnato il numero uno, assegna ai discendenti di un nodo interno un numero identificativo pari all'identificativo del nodo padre moltiplicato due per il discendente di sinistra e con l'aggiunta di una unità per il discendente di destra. Dunque, questo sistema di designazione risulta estremamente semplice ma efficace, infatti consente di risalire in maniera univoca alla posizione di un nodo nella struttura a partire dal numero ad esso assegnato.

Il punto cruciale è la scelta di queste suddivisioni, delle variabili che permettono di suddividere le osservazioni in due sottogruppi.

Differentemente dal modello di regressione ($y= ß_0 + ß_1x1 + ß_2x2 ...+ Ɛ$), nel modello considerato le covariate non compaiono come sono, ma **compaiono sotto forma di variabili dicotomiche** (family relation: Spose or not spose?...).

Infatti tutta la struttura ruota attorno a queste suddivisioni binarie, più genericamente rappresentano domande binarie del tipo: Xj , che è un generico predittore, appartiene ad A, sì o no? -Risposta dicotomica- Xj è quindi un fattore, una variabile qualitativa, come nel nostro esempio *family relationship*, che non assume un fattore numerico, le cui modalità sono degli attributi e di conseguenza l'insieme A non è altro che un insieme delle modalità del predittore.

Se invece Xj fosse una variabile numerica, il sottoinsieme A sarebbe definito su un intervallo di valori che la variabile numerica può assumere (età a scuola come nel nostro esempio).

Il numero di variabili di suddivisione che si possono generare partendo dalle variabili originarie e quindi dalle coviariate, è un numero che dipende dalla natura, dalla scala di

misura delle stesse, quindi è importante sapere se la variabile iniziale che viene presa in esame è in origine dicotomica, qualitativa ordinale, qualitativa nominale – non sussiste una relazione di ordine – oppure numerica. Quindi ogni covariata può generare un numero di split che dipende dalla sua scala di misura.

Questo specchietto riassume il numero di Split che un predittore può genare:

- Predittore **dicotomico:** (Maschio o Femmina). C'è **1** sola possibile suddivisione (o maschio o femmina).
- Predittore **numerico:** esistono *n* valori distinti che possono generare **n-1** suddivisioni binarie
- Predittore **qualitativo**, ma **c'è** una linearità d'**ordine** nelle modalità. Ad esempio i voti a scuola: sufficiente, buono, distinto, ottimo. Se si volesse splittare in due gruppi sulla base di questi livelli, si potrebbe dividere in un gruppo *sufficiente* e in un altro gruppo *buono-distinto-ottimo*, oppure *sufficiente-buono* in un unico gruppo e distinto-ottimo nell'altro, o ancora da un lato s-b-d e dall'altro ottimo. Quindi si avranno 3 possibili suddivisioni binarie a partire dai 4 livelli del predittore preso in considerazione; riassumendo: le modalità sarebbero **m** =4 e le suddivisioni sarebbero **m-1** =3
- Il caso più complicato dal punto di vista computazionale è quello dei fattori, delle variabili **qualitative** che **non hanno una relazione d'ordine** tra le modalità (*titolo di studio superiore: liceo classico, scientifico, linguistico*…etc sono attributi tra cui non esiste una relazione d'ordine). Si potrebbero raggruppare sotto forma di combinazioni lineari in molti modi; considerate m modalità della variabile, è possibile generare ben $\mathbf{2^{(m-1)} - 1}$ variabili di suddivisione, split (dunque con un grande peso computazionale).

Si pone dunque un primo problema nella costruzione dell'albero, ossia scegliere quale predittore o partizione binaria debba essere utilizzata per segmentare ciascun nodo; per quale principio i nodi terminali non vengano segmentati e come viene associato ciascun nodo terminale ad una classe di risposta.

Questi interrogativi delineano i punti cardine delle metodologie ad albero così classificati:

1. Un insieme di domande binarie risultanti dalla dicotomizzazione dei predittori;

2. Una regola, cosiddetto *criterio di split* che consenta ad ogni nodo interno di scegliere la domanda binaria in base alla quale suddividere il nodo;
3. Un criterio che consenta di dichiarare un nodo terminale;
4. Una regola di assegnazione di un valore di risposta a ciascun nodo terminale.

2.3.1. Regole di Splitting

Al fine di dividere il campione in sottogruppi e decidere quale valore soglia utilizzare sulle variabili, è necessaria una regola di splitting per decidere quale variabile, o combinazione di variabili, deve essere usata in un certo nodo.
La procedura di segmentazione binaria consiste in una partizione binaria ricorsiva di casi N in due sottogruppi disgiunti. Lo scopo della procedura è quello di definire una regola di classificazione / predizione sulla base di un set di apprendimento (chiamato anche training set), per i quali sono stati registrati valori di una variabile di risposta Y e di una serie di variabili esplicative XI, X2, ..., XK (sia numeriche che categoriche).

I dati sono partizionati scegliendo ad ogni passo una variabile e un punto di taglio lungo di essa in base a una bontà di misura divisa che consente di selezionare quella variabile (più precisamente una variabile di divisione) che genera i sottogruppi più omogenei rispetto alla variabile di risposta.

La procedura di partizionamento ricorsiva segue un algoritmo "divide and conquer", nel senso che in linea di principio l'algoritmo continua i nodi di partizionamento finché tutte le foglie contengono un singolo caso oi casi che appartengono alla stessa classe o che presentano lo stesso valore di risposta.

Di conseguenza, un ulteriore passo, la semplificazione dell'albero, è di solito effettuata per evitare l'overfitting e migliorare la comprensione dell'albero rimuovendo retrospettivamente alcuni dei rami (la cosiddetta potatura)

Una volta che la struttura ad albero è stata potata, ad ogni nodo terminale viene assegnata un'etichetta di classe o un valore di risposta

Riepilogando, i metodi basati sull'albero coinvolgono i seguenti passaggi:

1) crescita dell'albero

2) potatura dell'albero

3) assegnazione delle classi / valori di risposta ai nodi terminali.

Il punto è quello di scegliere tra le variabili di split la migliore, che garantisca la suddivisione delle osservazioni presenti nel nodo in sottogruppi il più possibili omogenei al loro interno ed eterogenei tra loro.

Occorre effettuare una sorta di **valutazione della bontà** degli *split*, dunque è necessaria una misura della loro qualità in termini di riduzione della eterogeneità, della variabilità della variabile di risposta in modo da poter scegliere lo *split* migliore (che suddivide le osservazioni in due sottogruppi che sono in assoluti i migliori, rispetto ad ogni altro possibile *split*, e il più omogenei all'interno); in letteratura si parla a tal proposito di ***Goodness of Split Measure.***[9]

Questo tipo di misure sono state proposte nell'ambito dell'intelligenza artificiale, e non nell'ambito della statistica, la prima in assoluto è stata proposta nel campo del *machine learning*, considerando una variabile di risposta di tipo numerico. Qualche anno dopo, uno studioso di intelligenza artificiale ha proposto una variabile di questo metodo per il caso in cui la variabile fosse stata non numerica e quantitativa, ma qualitativa e che denotava quindi l'appartenenza a dei gruppi.

In seguito queste procedure sono state esaminate anche in campo statistico e tal proposito gli autori del CART hanno sviluppato un framework metodologico, introducendo il concetto generico di ***Impurity*** (impurità) che ingloba in sé sia quello di variabilità associato al caso di una variabile numerica, e sia quello di eterogeneità o mutabilità associato al caso di una variabile di risposta categorico.

Considerando la fase di generazione dell'albero, si parte dalla totalità delle osservazioni appartenenti al training set e si procede con una divisione binaria in classi. Si deve però stabilire preliminarmente il metodo in base al quale effettuare gli *splits*, basato sulla definizione di **impurità.**

[9] Quinlan, J.R. (1986). Induction of decision tree, Machine Learning, 1, 86-106

L'impurità è una funzione della frazione delle osservazioni classificate in ciascuna classe. Ad esempio, se *N* è il numero delle osservazioni contenute nel training set, una prima divisione binaria delle unità porterebbe alla formazione di due classi contrassegnate con le etichette "1" e "2", ognuna delle quali possiede una determinata frazione della totalità delle osservazioni. Indicate con F_1 e F_2 tali frazioni, la funzione di impurità I(F_1 , F_2) può essere pensata come una funzione, avente valori in [0,1], che fornisce una misura di quanto le osservazioni siano correttamente distribuite nelle classi.

Essa è definita in modo tale che $I = 1$ se le osservazioni sono concentrate in una sola classe e $I = 0$ se esse sono divise in parti perfettamente uguali fra le due classi. Poiché l'obiettivo dell'algoritmo è quello di posizionare le osservazioni in classi prestabilite, esso cercherà effettuare questa classificazione nel modo più corretto possibile e non considererà quegli attributi o quei valori degli attributi che non portano ad una corretta classificazione delle osservazioni. Formalmente, l'algoritmo nelle varie fasi sceglierà degli attributi splitter tra quelli presenti nel training set e proverà diversi valori in modo che in ogni nodo venga minimizzata la funzione di impurità. Minimizzare tale funzione coincide col trovare nelle varie fasi gli attributi e il rispettivo valore soglia, che operano la classificazione più corretta possibile e che quindi dovrebbero dare in quella fase un'informazione maggiore.

Il procedimento è di tipo *iterativo* e si arresterà quando non sarà più possibile, manipolando la scelta degli attributi e/o il loro valore soglia, diminuire la funzione di impurità.

La **misura di impurità** per il generico nodo *t* è così definita:

$$\mathrm{I}(t) = \phi(p(1|t),\cdots, p(J|t))$$

Essa è quindi una funzione $\phi(.)$ non negativa di $\{p(j|t)\}$ tale che:

1. $\phi(.)$ è massima solo quando $p(j|t) = \frac{1}{J}$ per ogni *j*;
2. $\phi(.)$ è minima e pari a zero quando $\phi(1,0,\ldots,0) = \phi(0,1,0,\ldots,0) = \ldots = \phi(0,\ldots,0,1) = 0$;
3. $\phi(.)$è una funzione simmetrica di $p(1|t),\ldots, p(J|t)$;

Quindi l'impurità di un nodo è massima quando tutte le classi della variabile dipendente sono presenti nella stessa proporzione, mentre è minima quando il nodo contiene casi appartenenti ad un'unica classe.
I due metodi comunemente usati per commisurare l'impurità sono l'entropia e l'indice di Gini.

2.4. Misure di Impurità

2.4.1. Il tasso di errata Classificazione

Una prima misura di impurità ad un nodo *t è* costituita dalla stima della probabilità che i casi presenti nel nodo appartengano ad una classe diversa da quella cui è associata la probabilità più elevata; essendo tale probabilità stimata in generale a mezzo della proporzione di casi presenti nel nodo e appartenenti a ciascuna classe, tale misura altro non è che la stima cosiddetto di *risostituzione* del tasso di errata classificazione, definita pertanto come:

$i(t) = 1 - max_j\ p(j|t)$ dove j sta a designare una generica classe di risposta e $p(j|t) = n_j(t)/n(t)$ è la stima con $n_j(t)$ numero di casi del nodo t appartenenti alla classe j.

Si sceglierà quello *split s* che rende massima la seguente quantità:

$$\Delta i(s; t) = -\max j\ p(j|t) + \max j\ p(j!2t)p2t + \max j\ p(j|2t + 1)p2t+1$$

L'indice di errata classificazione è semplice ed immediato, anche se trattandosi di una funzione non lineare, presenta alcuni inconvenienti.

In generale si ha che $\max(aj) + \max(bj) \geq \max(aj + bj)$ da cui discende che $\Delta i(s; t) \geq 0$ per tutti gli *split*.

Nel caso in cui la variabile di risposta sia a due classi { 1, 2} può accedere che un nodo presenti una maggioranza di casi di classe 1 e che per tutte le possibili suddivisioni del nodo i due nodi discendenti presentino anch'essi una maggioranza di casi della stessa

classe; dunque si avrebbe $\Delta i(s; t) = 0$ per ogni *split*, e non sarebbe possibile identificare uno *split* migliore tra tutti.

In figura sono mostrati altri esempi di inconvenienti.

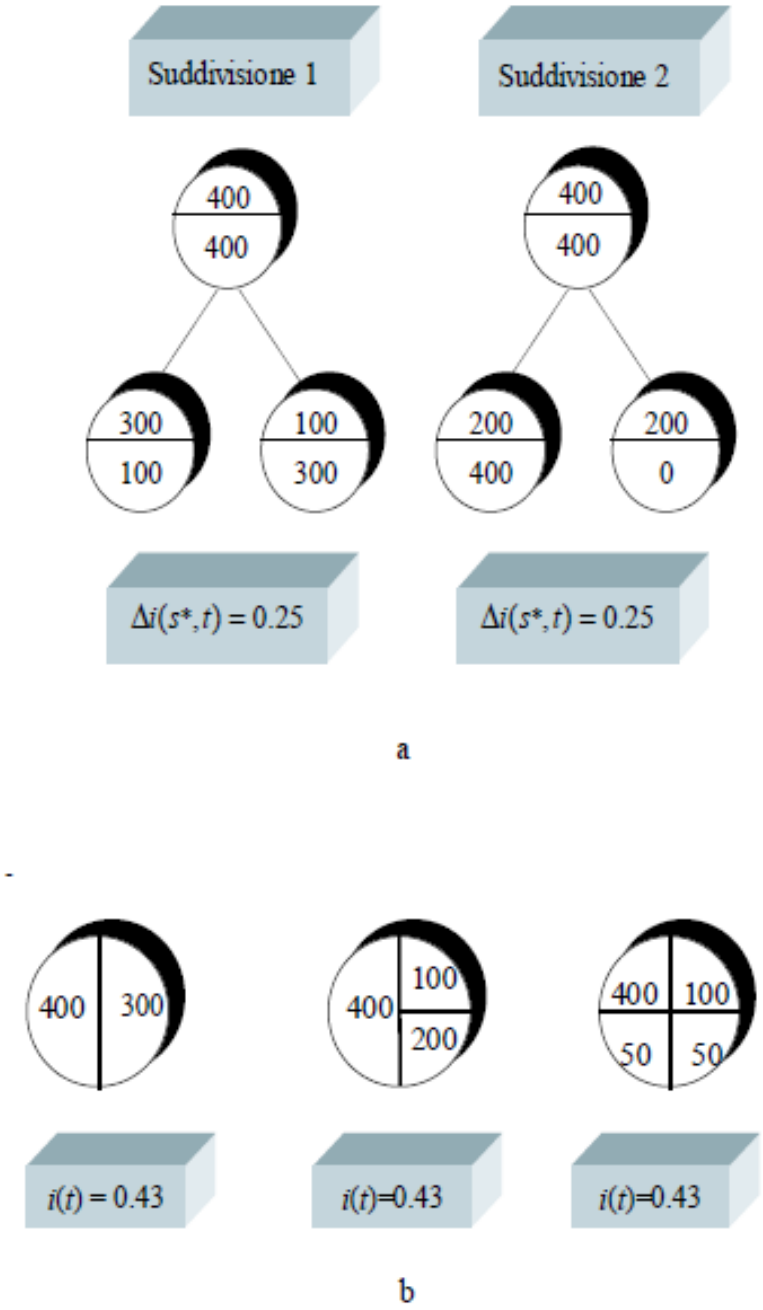

Figura 14- Inconvenienti del tasso di errata classificazione come misura di impurità- Non parametric regression and classification theroy and an application to web-marketing- Prof. C.Cappelli, 2016

2.4.2. L'indice di eterogeneità di Gini

L’indice di impurità che viene usato nella metodologia CART è quello di Gini, così definito al generico nodo *t:*

$$I(t) = \sum_{i \neq j} p(i|t)p(j|t) = 1 - \sum_j p^2(j|t)$$

Essendo $p(j|t) = 1 - p(i|t)$

Si nota come la prima parte dell’uguaglianza rappresenta la stima di errata classificazione di un’osservazione di classe *j* nella classe *i*, se l’assegnazione di un’unità del nodo *t* ad una particolare classe avvenga casualmente. La seconda parte è invece interpretabile in termini di varianza del nodo *t* qualora si codifichino, ad esempio, con “1” i casi di classe *j* appartenenti al nodo *t* e con “0” i casi di classe diversa.

Tale misura può essere interpretata come la stima della probabilità che un’osservazione scelta casualmente nel nodo *t* sia assegnata alla classe errata.

L'indice di Gini è una misura di impurità frequentemente impiegata per la creazione delle strutture ad albero essendo tra quelle implementate nei pacchetti statistici di segmentazione binaria. Anche tale indice non è esente da inconvenienti, in particolare esso tende a dare luogo a suddivisioni bilanciate dal punto di vista del numero di casi inviati dallo *split* nei due nodi figli ovvero ad evitare i cosiddetti *small splits*. In realtà nel caso di variabile criterio a due classi questa, che viene chiamata *anti-end cut pre- ference*, evitando nodi di piccola numerosità e quindi opponendosi all'arresto della procedura non costituisce un inconveniente perchè automaticamente uno *split* produrrà sottonodi che sono dominati da una delle due diverse classi.

Ciò potrebbe non accadere nel caso in cui la variabile di risposta sia multiclasse, di conseguanza l'indice del Gini fornirebbe sottonodi a fronte di una elevata purità e l’esclusività delle classi non sarebbe necessariamente perseguita come dovrebbe.

2.4.3. L’indice di entropia

Considerando un problema di classificazione con sole due classi, “1” e “2”, e sia *S* l’insieme delle osservazioni attraverso le quali si genera un albero di decisione.

Indicando con F_1 la frazione di esempi classificati con “1” e con F_2la frazione di esempi classificati con “2“, si definisce **entropia** di *S*, *H*(*S*), l’espressione:

$$\mathrm{H}\,(\mathrm{S}) = -\ F_1 \log_2 F_1 - \ F_2 \log_2 F_2$$

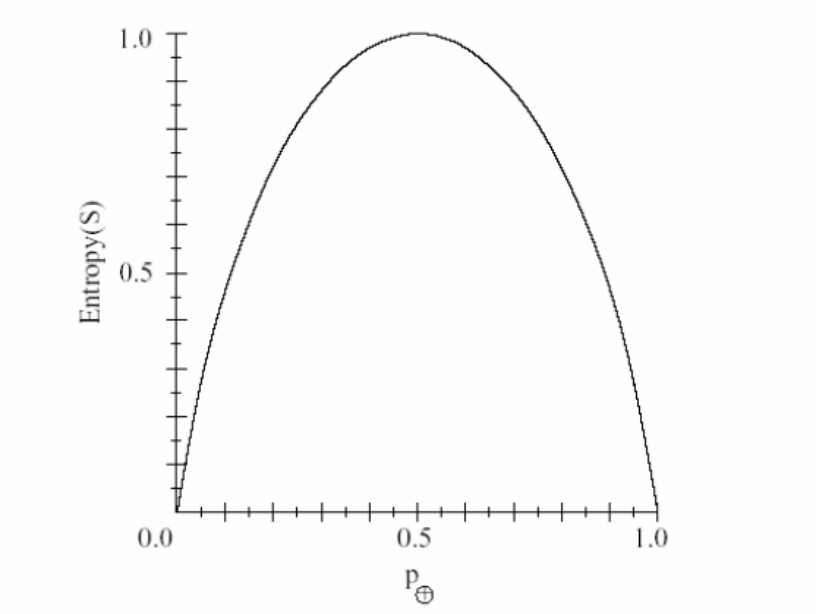

Figura 15- Indice di Entropia

La F*igura 14* riporta l'andamento dell'entropia al variare delle combinazioni delle frazioni F_1 e F_2. Dal grafico si evidenzia come $0 \leq H\ (S\ ) \leq 1$ ed in particolare, che il minimo valore dell'entropia, $H(S\ ) = 0$ è raggiunto nel caso in cui $F_1 = 1$ o $F_2 = 1$, quindi nei casi in cui la totalità degli esempi è classificata in una sola delle due classi.
Invece il valore massimo, $H(S) =1$ si ottiene nel caso in cui $F_1 = 0.5$ e conseguentemente $F_2 = 0.5$, ovvero nel caso in cui le osservazioni sono divise nelle due classi.
Generalmente, se le osservazioni sono classificate in *J* classi, l'entropia si definisce come:

$$\mathrm{H(S)} = -\sum_{j=1}^{J} F_j \log_2 F_j$$

L'entropia è una misura dell'impurità delle osservazioni che si considerano per la costruzione degli alberi di decisione; un valore elevato di entropia esprime la disomogeneità che caratterizza lo spazio dei dati, ovvero una maggiore difficoltà nell'assegnare ciascuna osservazione alla propria classe sulla base degli attributi che

caratterizzano la classe: più l'entropia è alta, più sarà difficile individuare gli attributi che effettivamente caratterizzano le classi.
Il caso delle sole classi "1" e "2" con un valore di entropia pari a 1, sarebbe il caso più difficile da considerare, se l'obiettivo è quello di individuare gli attributi che caratterizzano le due classi.
In generale, partendo da una situazione di massimo disordine in cui $H(S) = 1$ o da un qualunque valore elevato di entropia, una partizione dei dati effettuata rispetto ad un certo attributo ***A*** porterebbe ad un nuovo valore $H'(S)$ tale che risulta $H'(S) \leq H(S)$ e quindi ad una diminuzione di entropia. In tale ambito rientra il concetto di **information gain**, (guadagno di informazione) definito come la diminuzione di entropia che si ottiene partizionando i dati rispetto ad un certo attributo. Se indichiamo con $H(S)$ il valore iniziale di entropia e con (S, A) il valore dell'entropia dopo aver partizionato i dati in base all'attributo ***A***, l'information gain, che indicheremo con G, è data da:

$$G = H(S) - H(S, A)$$

Tale quantità è tanto maggiore quanto più elevata è la diminuzione di entropia dopo aver partizionato i dati con l'attributo ***A***. Dunque un criterio di scelta dei nodi di un eventuale albero di classificazione consiste nello scegliere di volta in volta l'attributo ***A*** che dà una maggiore diminuzione di entropia o che analogamente massimizza l'information gain.
L'information gain ha valori molto elevati in corrispondenza di attributi che sono fortemente informativi e che quindi aiutano ad identificare la classe di appartenenza delle osservazioni.
Spesso, però, più gli attributi sono informativi, più perdono di generalità; ad esempio, nel database di una compagnia telefonica, il campo codice fiscale è altamente informativo, ha dunque un alto valore di information gain, dal momento che identifica con certezza l'utente, ma non è per nulla generalizzabile. L'ideale è dunque individuare campi altamente informativi con un buon grado di generalizzazione.

2.5. Dichiarare un nodo terminale

2.5.1. Regole di arresto e dimensionamento retrospettivo

La semplicità e la facile intuizione di interpretazione sono dei vantaggi delle metodologie ad albero , anche se in caso la struttura risulti di grandi dimensioni e quindi contenga molte regole decisionali, caratterizzate dal concatenarsi di numerose condizioni, il vantaggio della semplicità interpretativa viene perso.

La dimensione eccessiva è dovuta al numero di nodi terminali o equivalentemente dal numero di suddivisioni rappresentati, quindi il problema è causato dal fatto che gli algoritmi di partizione ricorsiva che sono alla base della creazione delle strutture ad albero, seguono una strategia cosiddetta divide et impera.

Inizialmente,per ridurre la taglia della struttura ci si affidava alle cosiddette stopping rules, regole strutturate in modo tale da fermare la crescita dell'albero, e quindi dichiarare il corrispondente nodo terminale quando la riduzione dell'impurità conseguibile mediante la suddivisione del nodo stesso risulta inferiore ad una soglia prefissata, oppure fissare una numerosità minima dei nodi.

Naturalmente tali regole di arresto limitavano la possibilità di creare alberi articolati che di conseguenza non avrebbero potuto soddisfare la ricerca e l'osservazione di relazioni significative.

Gli autori del CART hanno proposto di costruire il cosiddetto albero totalmente espanso, o albero massimo indicato con Tmax, per poi rimuovere alcune branche in base ad un criterio cosiddetto di pruning.

2.5.2. Fase di assegnazione della classe di risposta al nodo terminale

Nella fase di assegnazione di una delle classi di risposta ai nodi terminali si possono presentare tre situazioni diverse in cui:

1. Il nodo contiene solo casi appartenenti alla stessa classe;
2. Il nodo contiene casi appartenenti a classi diverse, ma una di queste presenta una proporzione di casi maggiore delle altre;
3. Il nodo contiene casi appartenenti a classi diverse e nella stessa proporzione.

Nel primo caso l'assegnazione viene effettuata precisamente senza alcun tipo di indecisione, nel secondo caso il nodo viene assegnato alla classe che presenta più casi con una probabilità di assegnazione errata stimata pari a ($1 - max_j p\ (j|t)$). Nel terzo caso, invece, si presenta una situazione di massima incertezza, in quanto la probabilità elle classi risulta identica per ciascuna di esse, dunque si ricorre ad un tipo di assegnazione casuale salvo intervento del ricercatore che può effettuare un'assegnazione diversa sulla base delle sue conoscenze.
In generale, la regola di assegnazione è così riportata:
sia *HT* l'insieme dei nodi terminali *h* di un albero *T*, una regola di assegnazione è una funzione d (.) definita sull'insieme *HT* tale che:

$$\forall \boldsymbol{h} \in \boldsymbol{H_T} \rightarrow \boldsymbol{d(h)} = \boldsymbol{max_j p\ (j|h)}$$

2.6. Cenni sulla Regressione ad Albero

Nel caso in cui la variabile dipendente sia quantitativa invece che qualitativa, la metodologia CART genera alberi di regressione. L'approccio alla costruzione degli alberi di regressione è leggermente più semplice rispetto a quello per gli alberi di classificazione, dato che sia nella fase di costruzione che nella fase di pruning viene usata la stessa misura di impurità.
Si consideri una variabile dipendente *Y* che assume valore nel campo dei numeri reali e le p variabili esplicative X_1 ,... , X_p rilevate su *N* unità statistiche. La costruzione di un albero di regressione consiste nell'individuazione di una funzione (predittore o regola di previsione) $d(\mathbf{x})$ sullo spazio **X** delle variabili esplicative, che assuma valori reali.

Come negli alberi di classificazione, il predittore viene costruito suddividendo lo spazio **X** mediante split binari fino a raggiungere un insieme di nodi finali, con la logica della massimizzazione del decremento di impurità, impiegando come misura di impurità la devianza divisa per *n* osservazioni al generico nodo *t:*

$$R\ (t) = \frac{1}{n} \sum_{i=1}^{n\ (t)} (y_i - \bar{y}\ (t))^2$$

Dove $\bar{y}\ (t) = [\sum_i (y_i)] \ / \ n\ (t)$ rappresenta la media dei valori di risposta y_i al nodo t.
Ciascun possibile *split s* del nodo *t* induce un decremento nella misura di impurità *R*(*t*) definito come:

$$\Delta R\ (s,t) = R\ (t) - [\ R\ (2t) + R\ (2t + 1)]$$

Analogamente al caso della classificazione, sarà scelto per segmentare il nodo quello *split s* che massimizza la misura di impurità.
Essendo:

$$R\ (2t) = \frac{1}{n} \sum_{i=1}^{n(2t)} (yi - \bar{y}\ (2t))^2$$

E

$$R\ (2t+1) = \frac{1}{n} \sum_{i=1}^{n(2t+1)} (yi - \bar{y}\ (2t+1))^2$$

la quantità [*R*(2*t*)+*R*(2*t*+1)] può essere riguardata a meno del fattore costante *n*, come la devianza interna ai due gruppi indotti dalla suddivisione del nodo *t*, denotata, nuovamente con notazione mutuata dalla Analisi della Varianza, $WSS_{Y|s*}(t)$, dunque
il miglior *split* di un nodo è quello che rende massima la devianza tra i due gruppi indotti dalla suddivisione del nodo.
Si ricorre alla media aritmetica per la questione della assegnazione di un valore di risposta ai nodi terminali, essendo tale variabile numerica, e quindi ad ogni nodo terminale risulta associato il valor medio assunto dalla variabile di risposta per i casi caduti nel nodo.

2.7. L’accuratezza della struttura

Qualunque sia lo scopo della struttura ad albero ottenuta dall’analisi, quindi sia esplorativa dell’insieme di apprendimento o rispondente all’esigenza di fornire spiegazioni sulle osservazioni esaminate, si pone il problema di valutare l’accuratezza della struttura stessa.

Per accuratezza è inteso il grado di errore della struttura che per quanto riguarda la classificazione è misurato dal tasso di errata classificazione e per la regressione dalla somma dei quadrati (devianza).

Ai fini della valutazione di tale errore, la situazione ideale sarebbe quella in cui si disponesse di un insieme infinito di nuove osservazioni il cui valore di risposta fosse noto in modo tale da farle scivolare nella struttura per valutare quanto bene questa spieghi la variabile criterio e quindi quanto essa si discosti dal vero errore che verrebbe in tal modo non stimato ma determinato.

Sia nel caso della classificazione che in quello della regressione, non potendosi verificare questa situazione “utopica” occorre ricorrere ad una stima del vero errore prodotta dalla struttura.

A tal fine vi sono tre possibilità:

1. Stima di *risostituzione*;
2. Stima *test set*;
3. Stima *cross validation*[10]

2.7.1. La sima di Risostituzione

La stima di risostituzione si ottiene considerando le medesime osservazioni appartenenti all'insieme di apprendimento. Si avrà

$$R(T)= \sum_{h\in H_T} r(h)p(h)$$

[10] Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation, Journal of the American Statistical Association, 78, 316-330.

Con r(h) $=\frac{1}{n(h)}\sum_{i=1}^{n(h)} I(d(x_i \in h) \neq j_i)$,I(.) assume valore 1 se l'affermazione in parentesi è vera e quindi la classe assegnata all' osservazione *i*-esima (descritta dal vettore di predittori x_i) dalla regola *d* generata dall'albero differisce dalla vera classe, e valore 0 altrimenti.

La formula della stima di ricostituzione non è altro che il rapporto tra il numero di osservazioni dell'insieme di apprendimento mal classificate dall'albero *T* che indichiamo con *e*(*T*) ed il totale *n*.

Per il caso della regressione invece si calcola la seguente quantità:

$$R(T) = \sum_{h\in H_T} R(h) = \sum_{h\in H_T} \frac{TSS_y(h)}{n}$$

che, essendo $\sum_{h\in H_T} TSSy(h)$ la somma per tutti i nodi terminali della somma dei quadrati, divisa per *n* ne fornisce la media.

Il tasso di risostituzione, detto anche tasso apparente, essendo calcolato utilizzando i medesimi dati impiegati per costruire l'albero, dà una rappresentazione ottimistica della accuratezza μe ecco perché è assai poco utilizzato preferendosi il ricorso a stime ottenute con procedure che riducono il cosiddetto bias ottimistico..

2.7.2. La stima *test set*

La stima test set si fonda sulla casuale suddivisione dell'insieme di apprendimento *C* in due sottoinsiemi, C_1 e C_2 con $C_1 \cup C_2 = C \; e \; C_1 \cap C_2 = \emptyset$

L'insieme C_1 (generalmente pari al 70% dei casi) viene impiegato per dar forma alla struttura mentre l'insieme C_2 detto appunto *test set*, viene successivamente fatto scivolare nell'albero per valutare quanto accuratamente sia in grado di classificare o predire il valore di risposta dei casi in esso presenti.

Formalmente per la classificazione si calcola:

$$R^{ts}(T) == \sum_{h\in H_T} r^{ts}(h) p^{ts}(h)$$

dove con l'apice *ts* si è indicata la natura della stima ovvero l'appartenenza delle osservazioni considerate all'insieme *test*.

Normalmente si ricorre a tale metodo di stima quando *C* è di cardinalità elevata per non impoverire l'insieme di dati utilizzato per costruire l'albero.

.

2.7.3. La stima *cross validation*

La stima *cross validation*[11] viene utilizzata quando l'insieme di apprendimento non consente la distrazione di una parte delle osservazioni affinchè fungano da *test set.*

Questa stima permette di suddividere l'insieme di apprendimento in un numero *V* di sottoinsiemi di uguale numerosità C_1, ; C_v, …, C_V . Si costruiscono allora altrettante strutture ad albero *Tv* con le osservazioni rispettivamente di $C - C_1$, , $C - C_V$. Ciascuna di queste sarà poi validata con le osservazioni di volta in volta non impiegate, ottenendo così *V* stime *test set* R^{ts} (T_v) la cui media fornisce la stima cross validation associata alla struttura indotta impiegando tutte le osservazioni a disposizione.
Si avrà:

$$R^{cv}(T) = \frac{1}{V}\sum_{v=1}^{V} R^{ts}(T_v)$$

Si nota come ogni caso in C è utilizzato per costruire la struttura, usato una volta in un campione test, risulta dunque "parsimonioso" con i dati.

2.8. Caratteristiche della metodologia CART

I metodi ricorsivi di partizionamento sono stati sviluppati in numerosi e differenti campi di applicazione. Regioni di decisioni complesse, possono essere approssimate dall'unione di regioni più semplici. Uno dei più grandi passi in questo senso è stato fatto con

[11] Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation, Journal of the American Statistical Association, 78, 316-330.

l'introduzione della metodologia CART, un semplice metodo non parametrico per partizionare dati.

Una delle sue attrattive è la sua semplicità: effettua degli split binari su una singola variabile in modo ricorsivo; quindi classificare un campione può richiedere solo pochi semplici passi.

Nonostante la sua semplicità, è comunque in grado di ottenere risultati migliori di molti metodi tradizionali, su data set complessi, non lineari e composti da molte variabili.

La metodologia CART è stata inoltre valutata e testata a fondo sia da coloro che l'hanno sviluppata, sia da ricercatori nei più svariati campi di ricerca.

Comunque non si può affermare che le performance degli alberi di classificazione o regressione siano migliori in assoluto di ogni altro metodo, ma le loro buone performance, unite alla loro semplicità e consistenza, ne fanno un metodo largamente usato in molte discipline.

In particolar modo, la procedura CART è molto utilizzata per data set composti da variabili continue o discrete, sia nominali che ordinali, mescolate tra di loro, che di solito originano regioni di decisioni complesse. Alcuni aspetti tecnici della metodologia CART sono di primario interesse per gli statistici. Qui di seguito verranno elencati i principali sottolineando l'idea che ne è alla base.

Il CART è una procedura non parametrica, non richiede che le variabili siano selezionate in anticipo e i risultati sono invarianti rispetto a trasformazioni monotone delle variabili indipendenti. Ciò comporta che non c'è alcuna necessità in ambito applicativo di sperimentare trasformazioni monotone delle variabili indipendenti (logaritmi, radici quadrate, elevamento a potenza positiva, etc.). Nel CART tali variazioni non modificano i risultati a meno che lo split sia basato su combinazioni lineari di variabili. Il CART può utilizzare data set con struttura complessa, perché al contrario dei modelli parametrici, che sono stati messi a punto allo scopo di mettere in luce singole strutture dominanti nei dati, il CART è designato proprio alla scopo di trattare dati dotati di struttura multipla. È estremamente robusto all'effetto degli outliers, può utilizzare combinazioni lineari delle variabili per determinare gli split e può utilizzare congiuntamente variabili di tipo categorico e continue. In conseguenza di ciò il CART non richiede alcun pre-processamento dei dati. In particolare non è necessario rendere discrete le variabili quantitative.

Inoltre può mettere in luce dipendenze ed interazioni ed è utilizzabile se ci sono dati mancanti. Il CART produce alberi ottimali, utilizzando strumenti molto sofisticati per stabilire la loro accuratezza, può utilizzare la stessa variabile in punti differenti dell'albero e può essere utilizzato a supporto dei modelli parametrici convenzionali.
Infine il CART può essere utilmente impiegato per: l'analisi esplorativa dei dati, la selezione di variabili in ambito modellistico tradizionale, l'individuazione di interazioni rilevanti, il miglioramento di modelli, la semplificazione di processi di decisione, la selezione di predittori, la conversione di variabili quantitative in variabili qualitative, il trattamento di dati mancanti.

2.9. Implementazione in R

2.9.1. La funzione rpart ( )

In ambiente R sono disponibili una serie di pacchetti aggiuntivi utilizzati per generare di alberi di classificazione e regressione. I pacchetti computazionalmente più efficienti nonché i più recenti sono:

- `rpart.package`
- `tree.package`
- `mvpart.package`

Questi ovviamente includono, oltre ad un rilevante numero di funzioni ausiliarie, anche le corrispettive funzioni principali:

- `rpart()`
- `tree()`
- `mvpart()`

Le prime due funzioni sono molto simili e consentono entrambe di costruire sia alberi di classificazione che di regressione. In particolare si differenziano soprattutto per le diverse impostazioni di default, per il loro diverso grado di adattabilità e per il modo di trattare dati incompleti: la funzione `rpart()` segue infatti più fedelmente le idee di *Breiman* per la costruzione degli alberi mentre la funzione `tree()` quelle di *Venables* e *Ripley*.

La funzione `mvpart()` invece, la più recente, non è nient'altro che un aggiornamento della funzione `rpart()` in grado di costruire alberi di regressione multivariata cioè alberi in cui la grandezza da predire risulti essere un vettore di variabili continue invece che una semplice grandezza scalare. Perciò per quanto riguarda le generazione di alberi di classificazione la funzione `mvpart()` risulta praticamente del tutto equivalente alla funzione `rpart()`.
Di particolare importanza è la funzione `rpart()` per il semplice motivo che rappresenta l'unica funzione in grado di personalizzare il criterio di split da utilizzarsi durante la crescita dell'albero.

2.9.2. Input della funzione rpart( )

Una chiamata generica della funzione `rpart()` è della forma seguente[12]:

```
rpart(formula, data, weights, subset, method, model, x, y,
parms, cost, minsplit=20, minbucket=round(minsplit/3),
cp=0.01, maxcompete=4, maxsurrogate=5, usesurrogate=2,
xval=10, surrogatestyle=0, maxdepth=30, ...)
```

Le grandezze mostrate tra parentesi sono i nomi degli argomenti trattati ed i valori numerici eventualmente riportati sono le assegnazioni di default dei corrispettivi argomenti.

`formula e data`

[12]Terry Therneau, Beth Atkinson, Brian Ripley (June 29, 2015). Recursive Partitioning and Regression Trees.

Sono due argomenti strettamente legati che servono in primo luogo ad indicare in maniera sintetica quale sia la variabile da predire e quali siano invece le variabili predittive ed inoltre anche a indicare quali siano i valori che tali grandezze assumono all'interno del Learning Sample.
Se ad esempio si ha intenzione di costruire un albero di classificazione che predica una determinata classe J a partire dalle variabili X_1, X_2 e X_3 e si suppone che che i valori assunti da tali variabili all'interno del Learning Sample siano registrati nelle prime quattro colonne di un dataframe, si potrebbe procedere in uno dei seguenti modi:

1. chiamando direttamente i vettori in questione:

```
rpart(formula = LS[,1] ~ LS[,2] + LS[,3] + LS[,4])
```

2. oppure più comodamente nel caso in cui alle quattro colonne fossero stati assegnati rispettivamente i nomi `J, X1, X2, X3`:

```
rpart(formula = J ~ X1+ X2 + X3 , data = LS)
```

Si nota dunque come `formula` è un argomento obbligatorio, `data` invece è un argomento opzionale che indica in quale dataframe sono contenuti i dati del Learning sample.

`weights`

È anch'esso un argomento opzionale ed è utilizzabile ogniqualvolta si voglia costruire un albero di classificazione utilizzando costi di misclassificazione del tipo C(i) e non del tipo C(i|j). Esso potrà quindi essere un qualsiasi vettore indicante i diversi C(i), di lunghezza pari al numero delle classi di predizione ed avente tutte le componenti maggiori od uguali a zero. Il valore di default è ovviamente c (1, 1, … , 1).

`Subset`

È un argomento opzionale che indica- nel caso in cui i dati vengano presi da un dataframe - quali righe del dataframe, (quindi quali dati del Learning Sample) utilizzare per costruire l'albero. La scelta di default è quella di utilizzare tutti i dati disponibili nel dataframe.

Possiamo osservare che questo argomento verrà ovviamente utilizzato internamente da tutte quelle funzioni che necessitano di stime di crossvalidazione per le quali infatti è necessario estrarre dal Learning Sample una serie di subset.

`method`

insieme a `formula` è sicuramente l'argomento più importante della funzione `rpart()` in quanto indica la natura dell'albero che si vuole generare. Può assumere infatti quattro valori: `"anova"`, `"poisson"`, `"exp"` e `"class"` . I primi tre indicano la costruzione di alberi di regressione secondo tre diversi criteri, mentre l'ultimo, `"class"`, indica la costruzione di un albero di classificazione. Una volta fissato il valore di `method` l'argomento `parms` ci permette invece di assegnare un valore diverso da quello di default ad alcuni parametri del metodo prescelto.
In particolare nel caso in cui `method = "class"`, `parms` potrà essere una lista, di massimo tre elementi, contente l'assegnazione di una o più delle seguenti grandezze:

`"prior"`, `"lost"` e `"split"`.

`"prior"` è il vettore delle p(j) che di default assume il valore classico (N_A/N, N_B/N, … , N_J/N) con l'usuale significato dei simboli.
`"lost"` è invece la matrice dei costi di misclassificazione C(i|j) che di default sarà C(i|j) = 1- δ_{ij} .
`"split"` infine può assumere i valori `"gini"` o `"information"` e indica l'indice di impurezza da utilizzarsi per la scelta dello split ottimo e cioè o l'indice di Gini, che è anche il valore di default, oppure quello dell'Entropia.

Osserviamo che sebbene la funzione `rpart()` sia in grado di scegliere autonomamente il valore dell'argomento `method` a seconda delle caratteristiche del termine di sinistra dell'argomento formula (ad esempio infatti davanti ad una variabile J di tipo factor `rpart()` assegna a `method` il valore `"class"`) è comunque preferibile assegnare direttamente uno dei quattro valori sopra mostrati.

Non è raro infatti che le classi vengano indicizzate, spesso automaticamente, con numeri interi e che la funzione rpart interpreti erroneamente questi valori numerici come grandezze continue generando di conseguenza un albero di regressione invece che uno di classificazione.
Esiste un ulteriore possibile utilizzo dell'argomento `method` diverso quello di assegnargli uno dei quattro valori "`anova`", "`poisson`", "`exp`" e "`class`".
`method` può infatti anche essere utilizzato per implementare un albero di classificazione (o regressione) che utilizzi un criterio di split definito dall'utente.

`model`, `x` e `y`

Sono argomenti logici opzionali di scarso interesse teorico. Di default a tutti e tre viene assegnato il valore `FALSE`. Se ad uno o più di questi viene imposto il valore `TRUE` al risultato della funzione `rpart()` viene aggiunto rispettivamente una copia del modelframe, di `x` (matrice delle variabili predittive, nell'esempio iniziale la matrice costituita dalle colonne2, 3 e 4 del dataframe `LS`) e di `y` (vettore della variabile predetta, nell'esempio iniziale la prima colonna del dataframe `LS`).

La funzione `rpart()` individua l'albero ottimo utilizzando una strategia di tipo tradizionale cioè mediante una procedura suddivisa in due fasi successive:

- crescita secondo il massimo decremento di impurezza.
- potatura tramite analisi costo-complessità e stime di crossvalidazione.

Il processo di crescita e quello di potatura rispettivamente terminano ed iniziano dall'albero (Tmax), un albero solitamente di notevole complessità e caratterizzato dal fatto di essere il più piccolo albero avente tutte le foglie completamente pure relativamente ai dati contenuti nel Learning Sample.
Inoltre nella maggior parte dei casi l'albero ottimo ha una complessità significativamente inferiore a quella di T_{max} che in genere è dell'ordine del numero di dati del Learning Sample.

2.9.3. Output Della Funzione rpart()

La funzione `rpart()` restituisce in primo luogo come già detto l'albero ottimo che minimizza la stima di crossvalidazione
Per ogni nodo dell'albero vengono inoltre riportate una serie di grandezze in grado di descrivere brevemente il nodo in questione. In particolare, nel caso in cui `method = "class"`, ogni nodo sarà identificato da una stringa riportante le seguenti grandezze:

```
node), split, n, loss, yval, (yprob)
```

che rispettivamente mostrano: etichetta del nodo, split primario, casi di LS che passano per il nodo, classe predetta e stime delle probabilità di appartenza ad una classe, inoltre i nodi terminali sono individuati da un asterisco *.
Nel caso in cui invece si stia utilizzando un indice di impurezza personalizzato:

```
node), split, n, deviance, yval
```

che rispettivamente mostrano: etichetta del nodo, split primario, casi di LS che passano per il nodo, I(t) e classe predetta. Osserviamo che non troviamo riportato(yprob) cioè la stima delle probabilità di appartenza ad una classe in quanto, è dimostrato che gli alberi che meglio approssimano tali probabilità (ma non la classe di appartenenza) siano quelli cresciuti e potati secondo l'indice di Gini.
La funzione rpart(), oltre a quelle sopra citate, calcola anche molte altre grandezze tra cui:
`frame`
È un dataframe contenente tutte le informazioni relative ai nodi dell'albero generato. Ogni riga corrisponde ad un nodo dell'albero; le colonne invece riportano rispettivamente i valori delle grandezze `var, n, wt, dev, yval, complexity, ncompete, nsurrogate.`

`complexity`

indica il valore del parametro di complessità per il quale il nodo in questione collassa,

`ncompete`

indica il numero degli split competitori ed invece indica `nsurrogate` il numero degli split surrogati.

`Where`

Ogni componente di questo vettore indica il nodo nel quale termina la corsa del corrispondente dato del Learning Sample. La componente i-esima ci dice quindi in che nodo finisce il dato i-esimo; ogni nodo è individuato tramite la riga in cui esso si trova in `frame`.

`splits`

È un dataframe che riporta le caratteristiche di ogni split, competitori e surrogati inclusi. Tra queste ricordiamo: `count, improve e index`.
`count` indica per lo split primario e per i competitori il numero di dati del nodo mentre per quelli surrogati il numero di dati incastrati liberati dallo split surrogato.
`improve` per lo split primario e per i competitori indica il decremento di impurezza mentre per quelli surrogati la stima della probabilità di sovrapposizione.
Infine `index` che indica il punto di split.

`Cptable`

È un dataframe che riassume sinteticamente il processo di potatura mediante l'analisi costo complessità. Vengono infatti riportati per opportuni valori del parametro di complessità (α) alcune caratteristiche degli alberi corrispondenti tra cui: il numero di split e deviazione standard.

`x` e `y`

Sono rispettivamente la matrice delle variabili predittive ed il vettore delle classi di appartenza dei dati del Learning Sample. Sono presenti solo nel caso in cui venga esplicitamente richiesto nella chiamata della funzione `rpart()`.

3. : CASE STUDY PSA RETAIL ITALIA

3.1. Integrazione del web nel settore dell'Automotive: intro

Attualmente le economie stanno cambiando radicalmente, rivoluzione innescata dallo sviluppo dei mercati emergenti, dalla crescita accelerata delle nuove tecnologie, dalle politiche di sostenibilità. Engagement, integration, measurement rappresentano i tre pilastri della Digital Transformation avendo rivoluzionato molti settori tra cui quello dell'Automotive, basti pensare al cambiamento delle preferenze dei consumatori riguardo le auto di proprietà, la connettività e l'alimentazione.

Queste forze stanno dando origine a dirompenti tendenze nel settore automobilistico technology-driven: i servizi di mobilità, la guida assistita, il sistema elettrico-ibrido e connettività.

Lo studio da me affrontato propone un'ottica certamente non deterministica, ma offre un'esposizione del cambiamento nel processo d'acquisto nel settore, supportato da un'analisi del lead management all'interno di un dealer PSA, utilizzando la metodologia di classificazione ad albero.

Entro il 2020, la digitalizzazione in crescita e gli avanzamenti tecnologici avranno aumentato gli investimenti del settore automobilistico a 82 miliardi di dollari. L'industria automobilistica ha imparato rapidamente che bisogna soddisfare le richieste dei consumatori con un'esperienza digitalmente migliorata; i nuovi modelli di business potrebbero espandere il reddito automobilistico di circa il 30%, raggiungendo una soglia di $1,5 trilioni, tuttavia complessivamente le vendite di automobili hanno mostrato un

trend di crescita dal 2005 in poi, esclusi il biennio 2008-2009 (che ha risentito della crisi economica), con un incremento delle vendite dal 2007 al 2016 quasi del 29,2%.

Si evince dunque come il settore Automotive abbia subito molte variazioni nell'ultimo decennio, tanto che, dovendo rispondere alle attese del cliente, ogni sistema è stato integrato con una soluzione digitale.
Con le tecnologie digitali**,** i consumatori stanno cambiando il processo d'acquisto. Quando si recano presso le concessionarie automobilistiche, i clienti cercano informazioni specifiche, quelle che non trovano su Internet, e più che una presentazione dell'auto, hanno bisogno di consigli da parte di professionisti che abbiano una forte preparazione sul prodotto e sulla gestione dell'esperienza con il cliente**.**

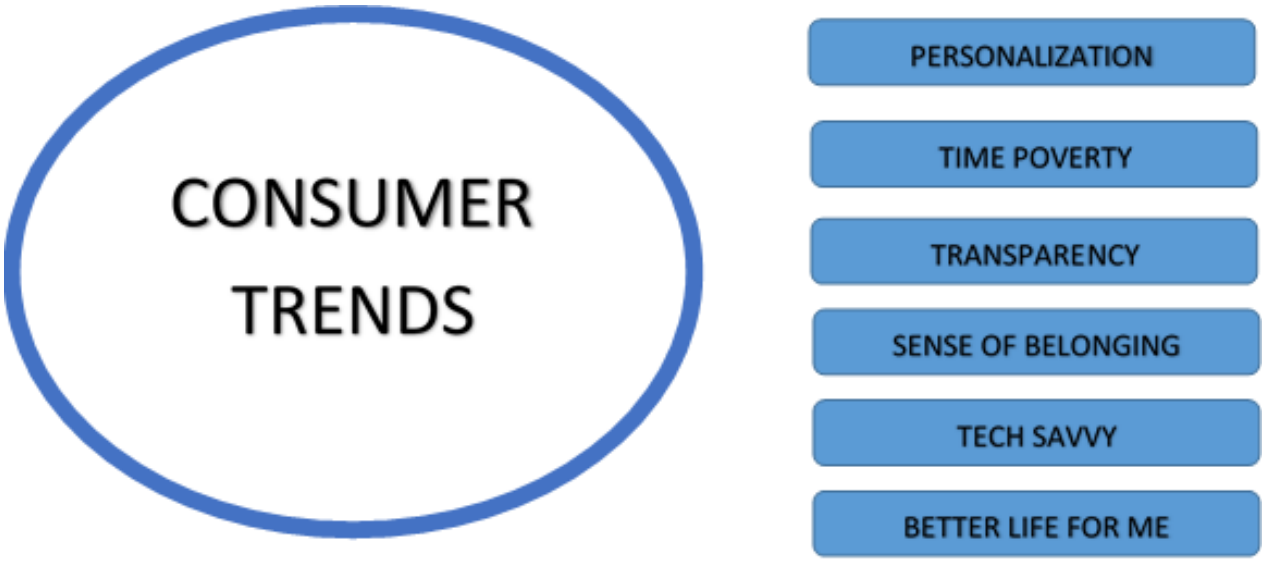

Figura 16- Consumer trends 2020- academy.quattroruote.it

Il concessionario del futuro è a portata di click, tanto che nel 2017 sempre più clienti prenotano un test drive o un tagliando tramite il sito web del concessionario o richiedono un preventivo dal proprio smartphone.

La Costumer Journey non inizia sulla soglia degli showrooom, ma sui touch point virtuali, quindi la prima sfida da affrontare consiste in una rivisitazione del processo di vendita. Di fondamentale importanza rappresenta il raggiungimento della consapevolezza che, dietro all'immaterialità di un contatto web, c'è sempre una persona e, quindi, un potenziale cliente. Sarebbe inammissibile lasciare al caso il primo approccio del consumatore sul digital, lasciandolo slegato dal resto del processo di vendita, quindi è necessario trovare un punto di collegamento tra la comunicazione sui canali tradizionali e quella sul web.

È quanto sostiene un'analisi realizzata da Accenture in Cina, Germania e Stati Uniti. Il commento all'analisi effettuata su tremila consumatori che hanno comprato un'auto nuova negli ultimi cinque anni, sottolinea come "i clienti nel settore Automotive, sempre più predisposti allo shopping online, si recano presso le concessionarie auto con sempre meno frequenza e solo per esigenze specifiche, spesso con l'obiettivo di finalizzare una decisione di acquisto già maturata sul web". (Mentuccia, 2016)

L'attuale esperienza d'acquisto legata a questi nuovi trend, però, sempre secondo lo studio di Accenture, non soddisferebbe appieno gli acquirenti per la scarsa integrazione tra l'esperienza maturata su internet e quella riscontrata nel punto vendita: in una scala da 1 a 4 punti, infatti, la media del giudizio espresso dagli intervistati sul livello di integrazione delle loro esperienze online e offline è stato di 2,32. A sorpresa, è emerso che chi è abituato a fare acquisti online visita la concessionaria prima della firma del contratto con una frequenza addirittura doppia rispetto ai consumatori più restii a fare shopping su internet. Per i primi, però, il tempo di tale visite sarebbe più breve perché vi entrerebbero con le idee già ben chiare.

Lo studio ha evidenziato anche che, oltre a esperti che possano dare consigli sul prodotto, molti degli intervistati vorrebbero avere a disposizione anche personale che possa rispondere a richieste di dettaglio e fornire raccomandazioni accurate relative al loro acquisto durante tutto il processo.

Si deduce quanto ci sia margine di miglioramento del settore nell'interazione con i clienti, utilizzando nuovi strumenti di CRM i quali non sono altro che applicativi in grado di raccogliere e strutturare dati e tecnologie per offrire esperienze di realtà aumentata e virtuale.

Durante il percorso di stage intrapreso in PSA Retail Italia, esaminando i dati del sito E-Dealer del marchio Peugeot, è risultato che il 70% degli utenti visita solitamente il sito naturalmente, l'8% da campagne e banner, 10% direct tramite Query Code che rimanda al sito E-Dealer, il 12% tramite il Referral; di conseguenza la decisione di non abbandonare alcun canale tradizionale aggiungendo alcune attività di marketing esperienziale con una serie di eventi sul territorio, cercando di ottenere la giusta combinazione tra quello che propone la casa madre a livello nazionale, e quello che comunicano i dealer a livello locale. L'obiettivo di tutta la comunicazione è diventato, quindi, quello di orientare il processo d'acquisto della nuova auto attraverso i presidi sul web, in modo da soddisfare le aspettative del cliente, in termini di risparmio di tempo facilitando la decisione d'acquisto.

Il sito web ha, pertanto, un ruolo ben preciso, infatti rappresenta una vera e trasparente guida all'acquisto. E' stato concepito una sorta di salone virtuale con una strategia in grado di riprodurre esattamente quella costruita per i clienti che visitano gli showroom dei dealer, motivo per cui oggi il cliente può sul sito configurare la vettura d'interesse e ottenere il prezzo finale della stessa.

Tutto è concettualmente pensato in modo da disegnare e orchestrare al meglio un Customer Journey ed una Customer Experience Omnicanale che siano quanto più coerenti e sinergici tra tutti i touchpoint, i protagonisti coinvolti, le azioni e i contenuti che li compongono, inoltre sono previsti percorsi personalizzati per ogni singolo interlocutore, con l'obiettivo di rispondere al meglio alle specifiche esigenze e al profilo di ciascuno.

Figura 17- Technology Advancements- academy.qattroruote.it

In tema di personalizzazione entra in gioco anche il canale social, un vero e proprio forum aperto h24 che dà la possibilità ai clienti di condividere esperienze con il marchio piuttosto che ricercare informazioni sui prodotti, dunque con la grande capacità non solo di comunicare i valori del brand ma soprattutto quella di gestire il rapporto stesso con il cliente.

I social media, rappresentano una grande opportunità per il settore in esame, basti pensare che nel 2017 sono 30mln gli utenti Web[13] (crescita dell'e-commerce del 20% rispetto al 2016) di cui 26mln da mobile, il 45% degli utenti social media hanno un'età >35 anni, molti dei quali hanno più di un profilo, inoltre il 61% degli acquirenti che si presentano in concessionaria si è informata precedentemente sul web.

Per aggiungere valore alle interazioni con i clienti sui social media, le aziende utilizzano diversi strumenti in grado di monitorare il Consumer Behaviour tramite opportune analisi.

[13] www.AudiWeb.com- il portale italiano che si occupa di stilare le statistiche sul web e dei suoi strumenti

I sistemi CRM sono utili ad alimentare i database sui clienti attraverso diversi canali o punti di contatto tra il cliente e l'azienda, includendo vari siti web della società e non, telefono, chat, direct mail, materiali di marketing e social media. I sistemi CRM possono anche dare informazioni dettagliate personali inerenti alla cronologia degli acquisti, le preferenze di acquisto e i vari contatti che ci sono stati tra l'azienda e il cliente.

Grazie a questi nuovi mezzi di comunicazione, strumenti fondamentali per la relazione con la clientela, si possono abbattere le barriere spazio temporali, riuscendo a creare una comunicazione simmetrica, interattiva e quasi istantanea.

La peculiarità dei software è quella di convogliare tutte le informazioni in un database CRM unico, in modo che il digital team possa accedervi facilmente. Le altre funzioni principali di questo software comprendono la registrazione di varie interazioni con i clienti (tramite e-mail, telefonate, social media o altri canali, a seconda delle capacità del sistema), automatizzando vari processi di workflow, quali attività, calendari e allarmi, e dando ai manager la possibilità di monitorare prestazioni e produttività in base alle informazioni registrate nel sistema.

Il sistema è fondamentale per snellire il lavoro del digital team, in quanto automatizza molti processi che richiederebbero del tempo, quali l'invio con cadenza periodica di mail e sms, dunque molte procedure che richiederebbero l'ausilio di un dipendente sono ridotte a pochi click.

L'avvento dei social media e la proliferazione dei dispositivi mobili ha portato i fornitori di CRM di aggiornare la propria offerta per includere le nuove caratteristiche che si rivolgono ai clienti che utilizzano queste tecnologie.

Cercando di integrare i dati di CRM sociale con altri dati dei clienti ottenuti dalle vendite o reparti di marketing al fine di ottenere una visione unica del cliente. Un altro modo in cui il CRM sociale è un valore aggiunto per le aziende e clienti dando vita ad una community, dove pubblicare recensioni di prodotti, interagendo con altri clienti per risolvere magari i problemi. Comunità di clienti possono fornire un servizio di basso livello per alcuni tipi di problemi e ridurre il numero di chiamate al call-center, possono anche aiutare le aziende, fornendo nuove idee per il futuro.

Al giorno d'oggi i clienti chiedono più vie di comunicazione con una società e si aspettano una perfetta interazione tra diversi canali, il più popolare tende ad essere la web chat,

applicazioni mobili e social media. La sfida principale di un sistema di CRM è fornire una customer experience coerente e affidabile.

3.2. Il CRM best-way per la comunicazione One To One e per la relazione con i clienti.

La comunicazione **One To One,** è **bidirezionale** e interattiva e gli attori coinvolti allo stesso tempo sono emittenti e riceventi.
Questo tipo di marketing è nato negli anni 90, in Giappone nelle catene *Just In Time*, dimostrando come miglior modo per centrare il target sarebbe stato quello di creare prodotti/servizi, in grado di rispecchiare realmente le esigenze del consumatore.
. L' approccio del CRM prevede delle fasi necessarie per l' implementazione:

- identificare gli utenti target;
- creare delle categorie di utenti;
- integrare la strumentazione per la comunicazione interattiva;
- personalizzare le strategie aziendali e le campagne di marketing.

La differenza sostanziale con il marketing di massa è proprio la possibilità di poter incrementare la penetrazione delle offerte al singolo utente con lo scopo di fidelizzarlo.
Il **contenuto** del messaggio deve essere mirato in modo da:

- accrescere il valore percepito dal cliente;
- rendere il valore percepito > del costo complessivo d' acquisto.

L'attività svolta può essere riassunta in tre step, pertanto avremo:

1. **Conoscenza e identificazione del target;**
2. **Interazione e personalizzazione;**
3. **Fidelizzazione.**

Conoscenza e identificazione del target: concretamente avviene tramite la generazione del ***lead,*** il cliente richiede un contatto con l' azienda utilizzando vari strumenti, il sito web, la pagina Facebook, testate automobilistiche online e altro.

Una volta intercettato il potenziale cliente ed essere riusciti a trasformarlo in Lead, inizia la fase più impegnativa, cioè il Lead Management. In questo step il ruolo del dealer è fondamentale, infatti è previsto per ciascun dealer un Business Development Center composto da persone formate per contattare in maniera efficace i lead generati e trasformarli in appuntamenti. Successivamente l' addetto avendo a disposizione delle informazioni, provvederà a classificarlo a seconda del prodotto d' interesse, della modalità di pagamento o del motivo della richiesta di contatto.

Conoscere il cliente significa ottenere quante più informazioni possano tornarci utili per monetizzare e concretizzare il contatto. Dopo aver ottenuto una base di dati adeguata sarà possibile la classificazione, **clienti TOP, clienti ad elevato potenziale, clienti medi e disinteressati.**

E' necessario aver un processo strutturato per la gestione di un Lead, perché si tratta di un potenziale cliente esattamente alla pari di chi varca la soglia del nostro showroom. Riscoprire, attraverso il digital la centralità del cliente e la cura che bisogna mettere in ogni momento di contatto con ciascun Lead, è senza alcun dubbio un'opportunità notevole.

Interazione e personalizzazione: successivamente si passa alla progettazione di attività di comunicazione coerenti con ciascun "utente nel database", si pianificano mezzi, strumenti e contenuti della comunicazione. Definendo poi l' offerta personalizzata, il prodotto sulle specifiche esigenze del cliente.

Fidelizzazione: in questo step si mira a creare una relazione fiduciaria, la **customer loyality.** Il portafoglio clienti aumenta la redditività in maniera proporzionale e costante. Il cliente viene accattivato tramite operazioni di Trading-up, selling-up e cross selling (ad esempio il prolungamento di alcuni benefit per i clienti più profittevoli, riduzione di alcuni costi e fidelizzazione in senso stretto). Queste azioni legate tra di loro forniscono anche FeedBack fondamentali per il miglioramento continuo e progressivo del sistema CRM, essendo un processo che si sviluppa attraverso la rotazione delle informazioni già possedute e non dall' azienda.

Questo continuo circolo di informazioni e di feedback, determina lo sviluppo di una nuova conoscenza, di delle nuove linee guida che servono per migliorare progressivamente le azioni di marketing. L'azienda nel tempo acquisirà un Know-How relazionale che consente di ottimizzare e perfezionare i rapporti con la clientela.

3.3. L' infrastruttura informatica alla base del CRM

Le aziende che si affacciano al mondo del CRM devono svolgere dei cambiamenti significativi nell'assetto tecnologico e organizzativo.

Per realizzare il processo relazionale, è necessaria una piattaforma informatica dotata di hardware e software, consentendo di mettere in moto le diverse fasi del sistema in tempi rapidi e simultanei.

Possiamo suddividere il tutto in due grandi insiemi:

Il Management (CRM) è un processo continuo effettuato dalle imprese per creare costantemente nuove opportunità e mantenere relazioni profittevoli con i propri clienti. L'industria automobilistica si impegna costantemente per implementare nuove strategie da integrare con le proprie strategie industriali.

CRM Operativo si suddivide in:

- Gestione della relazione: Si cerca di attivare una molteplicità di strumenti per gestire in modo profittevole la relazione con i clienti, i diversi punti di contatto e le interazioni con i clienti.
- Personalizzazione: Si sviluppano prodotti e servizi ad hoc per ogni singolo cliente.

Fidelizzazione della clientela: Si sviluppano relazioni fiduciarie e di customer satisfaction, aumentando la redditività di impresa di lungo periodo.

- Feed-back: Si attua un processo di feedback per comprendere i nuovi bisogni dei consumatori e modificare conseguentemente l'offerta.

Analytical CRM si riferisce all'attività di CRM in cui si sfruttano i dati dei clienti per aumentare il valore apportato dal cliente per l'azienda. i dati dei clienti solitamente consistono in: nome, indirizzo, telefono, occupazione e reddito.

CRM strategico si riferisce al nuovo modo di creare vantaggi competitivi.

Con l'implementazione di CRM strategico, le aziende realizzano un valore aggiunto che può produrre un elevato livello di soddisfazione del cliente.

Il top management guida il customer- centric a tutti i livelli dell'organizzazione, così , focalizzandosi sul cliente, facilita l'apprendimento dei bisogni dello stesso e fornisce un valore aggiunto tramite la personalizzazione dell'offerta, mirando alla fidelizzazione.

Vanno inoltre classificati due tipologie di sistemi di CRM:

- **Client server**, legato strettamente ad un hardware e dei database offline;
- **E-CRM**, tutto ciò che è legato al web.

Entrambi i sistemi permettono la gestione di moli enormi di dati, provenienti da diverse fonti di contatto, tramite questi dati si passa alla segmentazione comportamentale del cliente, in base ad indici e probabilità di ritorno del contatto. Entrambe le strutture si basano su operazioni di front e back office, i**l CRM analitico è** costituito da hw/sw di supporto al back-office aziendale per svolgere le attività di identificazione, conoscenza e classificazione.

Il contatto *Lead,* arriva da una fonte esterna e successivamente passa nei compiti del front office.

Il CRM gestionale da forma al sistema ed è composto da tutte le attività applicative dedite all' interazione personalizzata con i diversi target.

La struttura informatica è di fondamentale importanza, sbagliare il sistema acquistato reca danni dovuti alla scarsa integrazione dei sistemi informatici ereditati dal passato (sistemi legacy) e alla mancanza di scalabilità, vale a dire all'utilizzo di software non in grado di sopportare la frequenza d'uso e il numero di utenti.

Rischi di questa natura sono, in ogni caso, evitabili attraverso un'analisi ex ante delle reali esigenze di CRM dell'impresa partendo dalla fattibilità dell'integrazione dei sistemi di back-office e front-office fino ad arrivare all'esame dei bisogni di accessibilità dei contenuti sui clienti.

Tale analisi dovrebbe condurre all'adozione di un'infrastruttura tecnologica di CRM affidabile e flessibile, che garantisca velocità di interazione con i clienti e sostenga, parallelamente, la comunicazione su tutti i media disponibili, in un'ottica multicanale.

Un ulteriore pericolo è relativo al sovradimensionamento del sistema, vale a dire a soluzioni applicative eccedenti le necessità aziendali. Per evitare ciò può essere efficace

ricorrere a soluzioni in outsourcing le quali, però, hanno rischi potenziali molto più grandi rispetto a quelle in-house.
La scelta di esternalizzare, infatti, comporta un grado di controllo sui dati dei propri clienti piuttosto basso, creando significative preoccupazioni inerenti la sicurezza e la duplicazione (back-up) di tali dati.

3.4. Analisi dei Lead con la metodologia ad Albero

3.4.1. Premessa

Durante il mio stage presso PSA Retail Roma, concessionario direct Peugeot Automobili Italia S.p.a. ho avuto modo di utilizzare un applicativo dedito al trattamento dei Leads, propriamente chiamati con l'acronimo GDO che indica la Gestione Delle Opportunità, che consiste in un programma creato appositamente da BERI S.p.a. per il marchio Peugeot. Tramite le strategie di Web Marketing, le aziende si impegnano di presidiare i Social e il Web dando la possibilità agli utenti di entrare direttamente in contatto con l'azienda tramite banner che portano alla landing page come mostrato nella *Figura 18*.

Figura 18- Banner e Skin crete in occasione degli Internezionali di Tennis BNL, presidio Siti

Figura 19- Landing Page Sito Web, presidio sito

Inoltre, da Gennaio 2016 Peugeot Automobili Italia ha avviato un processo di gestione delle Pagine Facebook di tutti i concessionari d'Italia, co-amministrando tutte le Fan Page e alimentando il database del CRM, tramite applicazioni presenti sui social che permettono di richiedere un preventivo, informazioni, test-drive, brochure, richieste anch'esse che "atterrano" sulla Landing Page.

Successivamente l'utente è portato a compilare un format presente sulla pagina indicando nei campi obbligatori alcune informazioni utili all'azienda per comprendere il tipo di richiesta e i contatti dello stesso che si presenterà tramite la piattaforma di CRM sotto forma di GDO, con lo scopo di fornire un supporto o un primo contatto per poi avanzare il rapporto verso l'acquisto.

La piattaforma, che prende il nome di START, ha lo scopo di:

- Fornire una visione a 360° dei contatti/reports
- Avere un punto d'entrata comune tra gli attori coinvolti
- Semplificare la gestione dei leads
- Migliorare l'impiego degli strumenti a disposizione della Rete.

3.4.2 Il Dataset

Il Dataset considerato in questa analisi è costituito da tutti gli utenti trasformati in GDO, dunque tutte le richieste di preventivo, di test drive o di informazioni effettuate tramite web (landing page o Facebook) con la relativa profilazione dell'interessato e la

conversione delle richieste in eventuali appuntamenti, prevenivi effettuati in sede e relativi contratti.

Le variabili prese in esame si riassumono come segue:

1) Città (binaria con lee modalità: Roma, fuori Roma,)
2) Nome e Cognome dell'utente
3) Intervallo di tempo tra contatto e appuntamento in filiale (in giorni)
4) Tipo di contatto (Richiesta informazioni, richiesta preventivo, richiesta test drive)
5) Fonte della richiesta (Sito web o Facebook)
6) Preventivo (si/ no)
7) Contratto (si/ no→ binaria, come variabile di risposta).

Il periodo d'interesse corrisponde al primo trimestre del 2017 (Gennaio- Febbraio-Marzo), in cui ho registrato in totale 341 richieste totali di cui 7 richieste di Informazioni, 289 richieste di preventivo e 45 richieste di Test-drive; 131 richieste in totale provengono da Facebook, 210 dal Sito Web come si evince dal grafico.

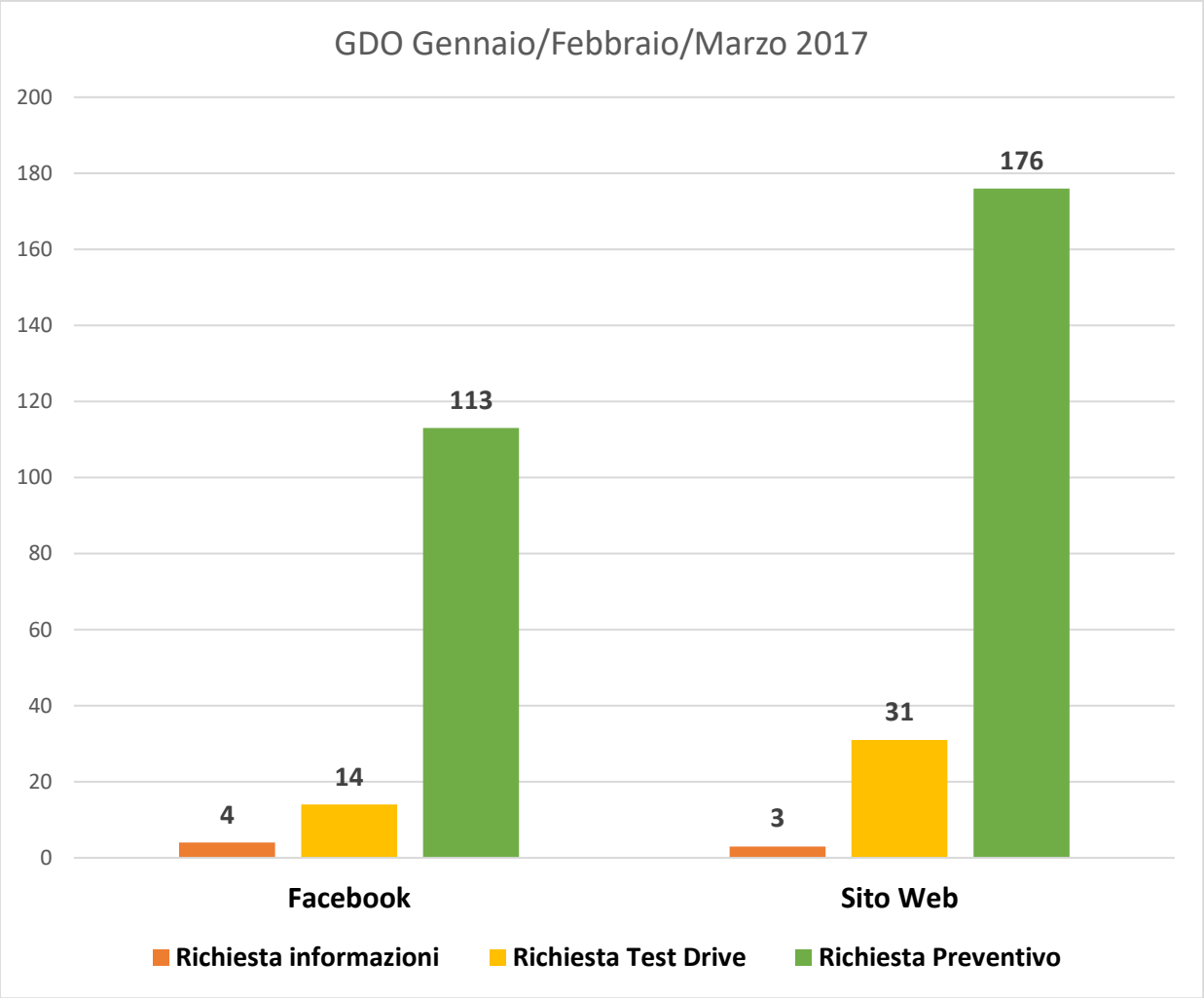

Figura 20- Richieste Online primo trimestre 2017 Peugeot filiale di Roma

Di questi 341 Contatti totali, tramite un'opportuna attività da parte del Business Development Center composto da persone formate sono stati contattati efficacemente i lead generati e trasformati in 146 appuntamenti in filiale, di cui 94 Preventivi con Test Drive effettuati in sede e una conversione in contratti del 49.47% (46 contratti in sede) come si evince nella *Figura 20*.

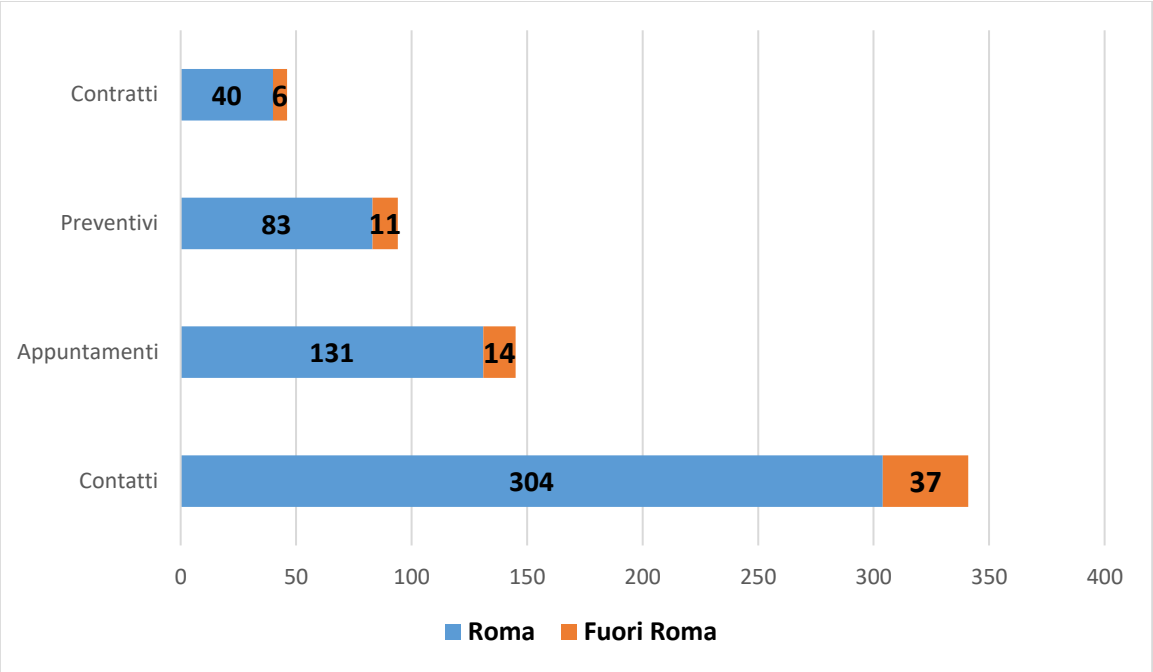

Figura 21- Conversione in contratti Peugeot Filiale di Roma

3.4.3. L'analisi in RStudio

L'analisi è effettuata per mezzo del software RStudio considerando il data set precedentemente esposto, quindi la variabile di risposta (y) è rappresentata da Contratto, che indica i Contratti effettuati nel primo trimestre del 2017, in seguito ad una richiesta online effettuata sul sito Web o Facebook. Le variabili esplicative sono: Città (Roma o Fuori Roma), Fonte (Sito web o Facebook e relativa compilazione del format), Preventivo (dicotomica→ sì/no), Tempo (tempo in giorni intercorsi tra la richiesta online e quindi il primo contatto e l'appuntamento in filiale).

Il primo step, dopo aver trasformato il file Excel in formato csv viene caricato in RStudio.

```
dataset.Micaela<-
read.csv("C:/rimerfimo/Downloads/statistica/micaela
brindisi /dataset Micaela Brindisi.csv")

View(dataset.Micaela)
```

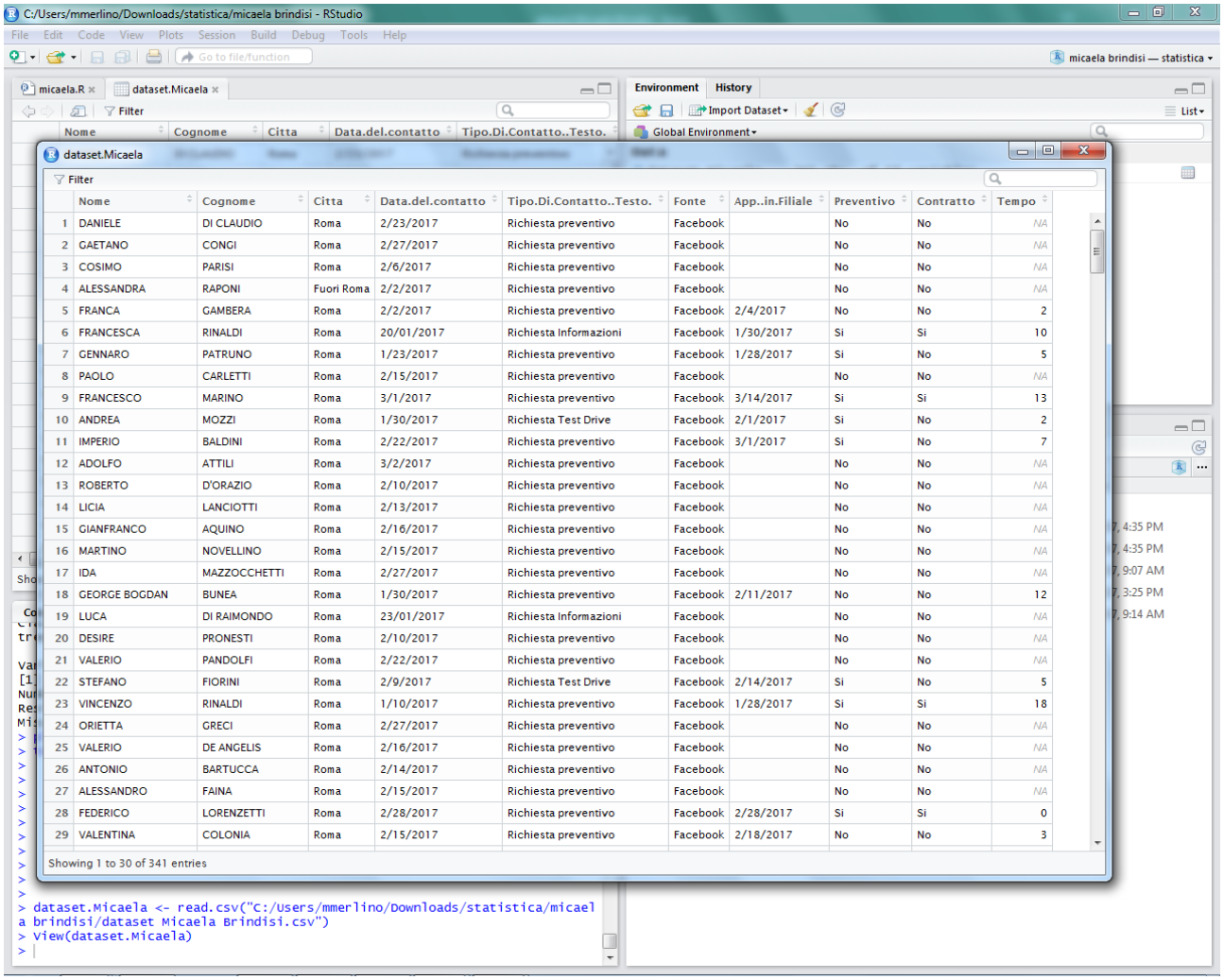

Figura 22- Dataset in RStudio

Carico il pacchetto ‘rpart’ e costruisco l’albero di classificazione:

```
library(rpart)

fit <- rpart(Contratto ~ Citta + Fonte + Preventivo+ Tempo,
method="class", data=dataset.Micaela)

printcp(fit)

fit
```

Di cui l'output:

```
Classification tree:
rpart(formula = Contratto ~ Citta + Fonte + Preventivo + Tempo,
    data = dataset.Micaela, method = "class")

Variables actually used in tree construction:
[1] Fonte      Preventivo Tempo

Root node error: 46/341 = 0.1349

n= 341

        CP nsplit rel error xerror    xstd
1 0.065217      0   1.00000 1.0000 0.13714
2 0.021739      2   0.86957 1.2391 0.14978
3 0.010000      7   0.76087 1.1304 0.14432


Variable importance
Preventivo      Tempo      Fonte
        87         10          3


node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 341 46 No (0.8651026 0.1348974)
   2) Preventivo=No 247  0 No (1.0000000 0.0000000) *
   3) Preventivo=Si 94 46 No (0.5106383 0.4893617)
     6) Fonte=Sito web 56 24 No (0.5714286 0.4285714)
      12) Tempo< 4.5 39 15 No (0.6153846 0.3846154)
        24) Tempo>=2.5 14  3 No (0.7857143 0.2142857) *
        25) Tempo< 2.5 25 12 No (0.5200000 0.4800000)
          50) Tempo< 1.5 13  5 No (0.6153846 0.3846154) *
          51) Tempo>=1.5 12  5 Si (0.4166667 0.5833333) *
      13) Tempo>=4.5 17  8 Si (0.4705882 0.5294118) *
     7) Fonte=Facebook 38 16 Si (0.4210526 0.5789474)
      14) Tempo< 7.5 29 14 Si (0.4827586 0.5172414)
        28) Tempo>=4.5 8  3 No (0.6250000 0.3750000) *
        29) Tempo< 4.5 21  9 Si (0.4285714 0.5714286) *
      15) Tempo>=7.5 9  2 Si (0.2222222 0.7777778) *
```

Nella costruzione dell'albero sono state usate le variabili 'preventivo', 'tempo' (il periodo di tempo intercorso tra contatto e appuntamento) e 'fonte'; non è stata usata la variabile 'città'.

Abbiamo 8 nodi terminali, identificati dall'asterisco; in ognuno di essi la ripartizione della relativa modalità di contratto è espressa come numerosità e come probabilità tra parentesi.

Si nota nel primo nodo (nodo radice), in corrispondenza di preventivo=no, abbiamo 247 soggetti per cui contratto=no e 0 soggetti per cui contratto=sì, una classificazione quindi perfetta a cui corrisponde, nella parentesi, 1 per il no, cioè 100% e 0 per il sì, 0%.

Da questa prima analisi risulta evidente come la variabile, nell'ambito delle variabili esplicative, che maggiormente suddivide le osservazioni in due sottogruppi omogenei sia "Preventivo".

Infatti nella costruzione dell'albero la variabile "Preventivo" ha un'importanza dell'87%, la variabile "tempo" del 10% e la variabile "fonte" del 3%.

Il misclassification rate (o prediction error) può essere calcolato come = Root node error * rel error * 100%

Nel nostro caso misclassification rate= 0.1349*0.76087*100=10.26%, presenta un valore molto basso.

```
pred <- predict(fit, type="class")
table(dataset.Micaela$Contratto,pred)
```

```
pred
     No  Si
  No 271  24
  Si  11  35
```

Tramite la funzione `predict` possiamo confrontare la classificazione reale con quella stimata dall'albero; nella tabella leggiamo per riga la classificazione reale e per colonna quella inferita dalle funzioni lineari utilizzate.

Sulla diagonale principale si hanno i casi in cui le classificazioni coincidono. Dunque abbiamo 24+11 casi di classificazione errata corrispondenti ad un error rate di 35/341=10,26%, che coincide con il misclassification rate calcolato in precedenza.

A questo punto per avere una migliore rappresentazione grafica dell'albero costruito utilizziamo il pacchetto **rpart.plot**, ottenendo il grafico presente in *Figura 22*.

```
library(rpart.plot)
rpart.plot(fit,main="Classification Tree for Contratto")
```

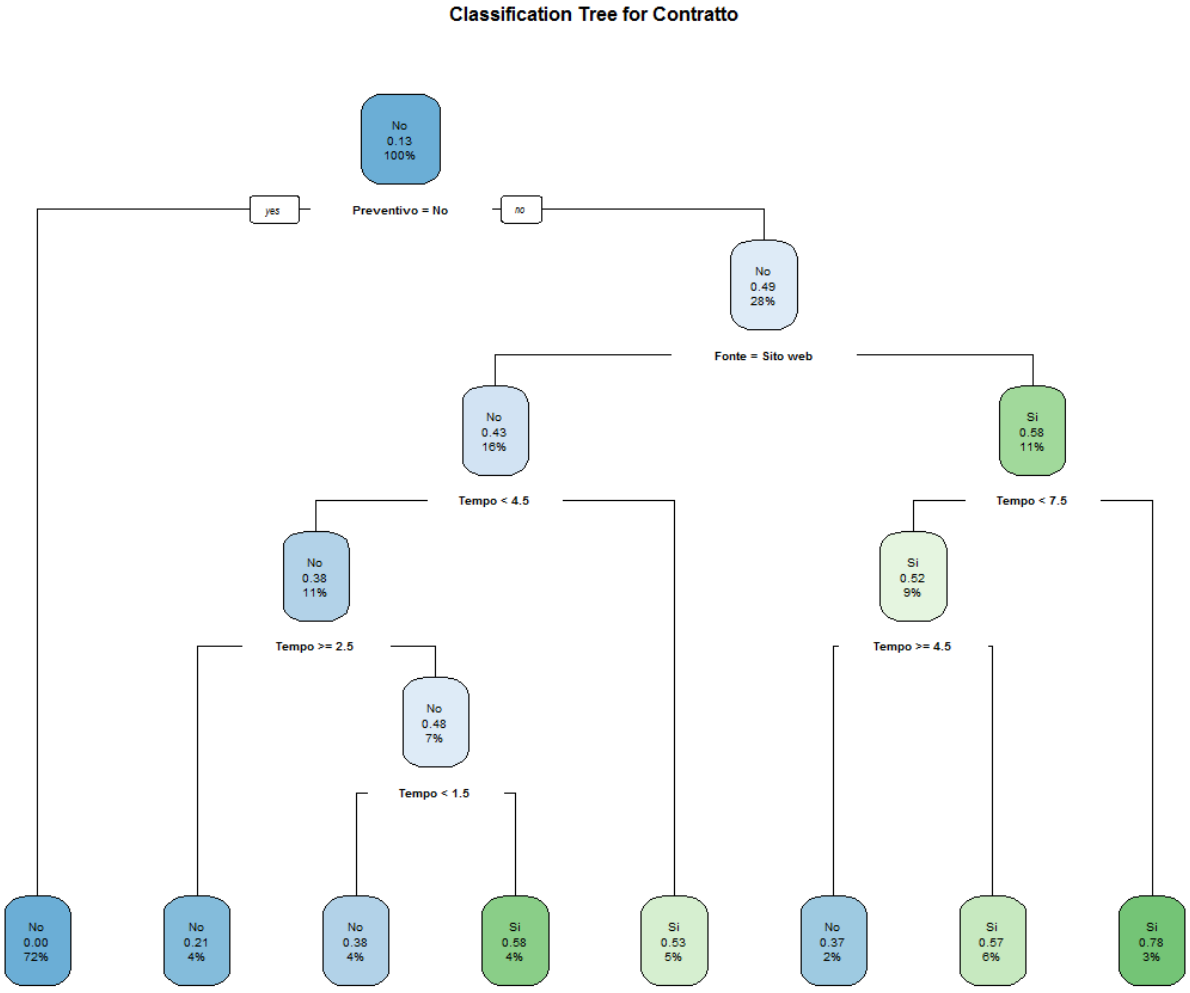

Figura 23- Classification tree for Contratto

In figura si notano gli 8 nodi terminali nell'ultima riga, insieme all'errore di classificazione relativo alla modalità di contratto evidenziata e al percentuale di soggetti rispetto al totale del dataset (ad es. nel secondo nodo terminale c'è il 4% di soggetti totali e classificandoli come contratto=no il tasso di errore è pari a 0,21, cioè al 21%).

Alternativamente possiamo ottenere il grafico con la funzione:

```
prp(fit, main="Classification Tree for Contratto",faclen = 0, cex
= 0.7, type=2, extra=101,box.palette="auto")
```

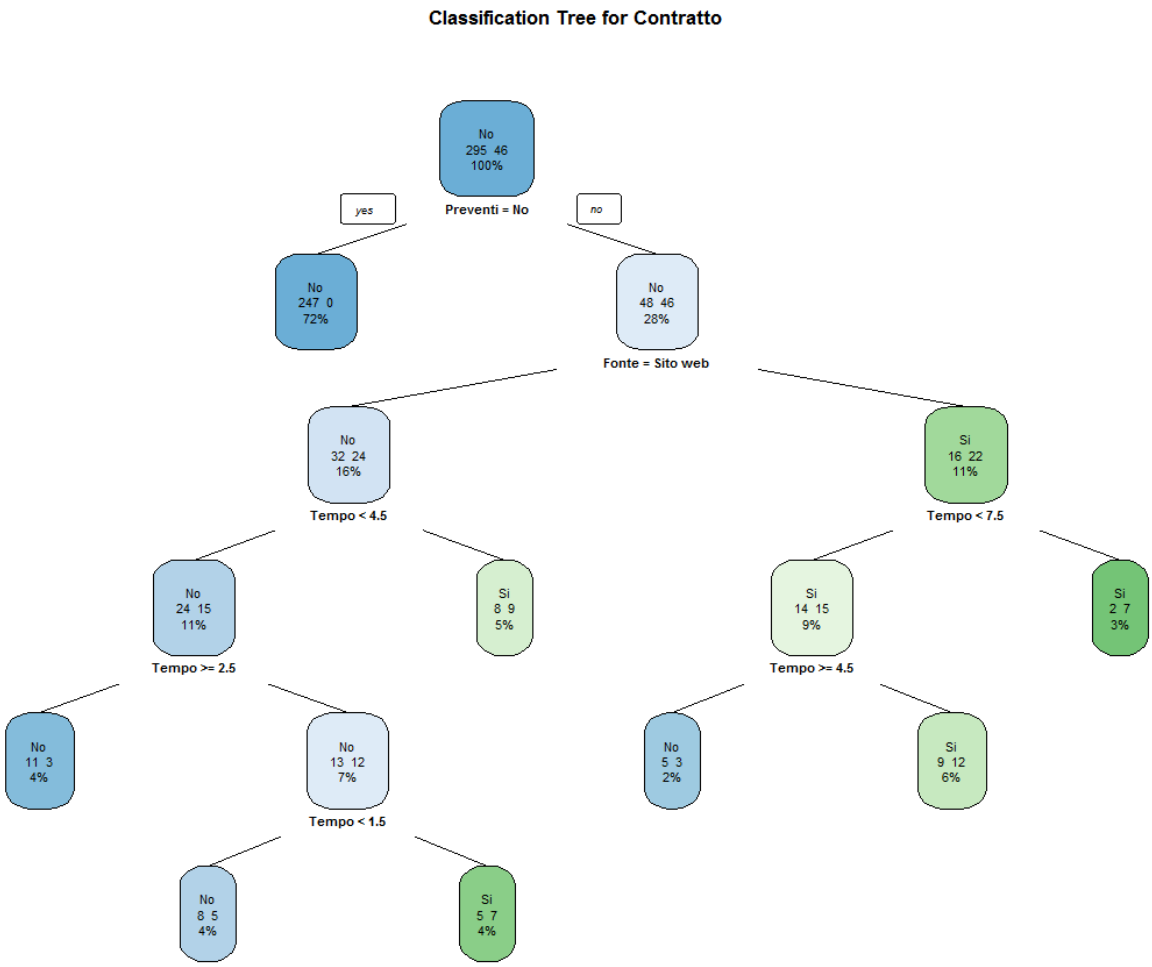

Figura 24- Classification Tree

I risultati sono ovviamente gli stessi, vediamo i nodi terminali non più allineati in fondo ma correttamente in base al numero delle intersezioni e inoltre al posto dell'error rate abbiamo le numerosità assolute di sì/no per la modalità relativa (ad es. il nodo visto in precedenza, a cui si arriva con preventivo=no, fonte=sito web, tempo < 4.5, tempo >=2,5, ha ovviamente la stessa percentuale di soggetti totali, 4%, e classificandolo come contratto=no abbiamo 11 soggetti classificati correttamente e 3 classificati non

correttamente, a cui corrisponde un error rate pari a 3/14=0.21, come nel grafico precedente.

In conclusione, avendo considerato la variabile binaria "contratto" come variabile dipendente, dall'analisi effettuata con le metodologie ad albero si nota come le variabili che influenzano l'esito del contratto siano il preventivo, la fonte di provenienza della richiesta e il tempo intercorso tra il primo contatto dell'utente e l'appuntamento in filiale del GDO che automaticamente è diventato un vero e proprio Lead.

CONCLUSIONI

Gli alberi di classificazione rappresentano una metodologia che ha l'obiettivo di ottenere una segmentazione gerarchica di un insieme di unità statistiche mediante l'individuazione di "percorsi" che sfruttano la relazione esistente tra una classe di appartenenza e le variabili rilevate per ciascuna unità.

Si parla di albero **binario**, poiché ogni nodo viene diviso in due parti, mediante lo Split.

Tali alberi sono numerati con un sistema di numerazione di tipo binario, utilizzato perché si tratta di ***numerazione univoca***, che ci consente di risalire alla posizione del nodo in una struttura ad albero, qualunque sia la forma della struttura

Se si adottasse una numerazione di tipo *consequenziale* si avrebbe una situazione in cui la numerazione del nodo dipenderebbe dalla posizione del nodo in quella particolare struttura, quindi i nodi sarebbero numerati in modo successivo. Di conseguenza, un nodo non sarebbe univocamente identificato nella struttura e dipenderebbe dalla forma del grafo.

Grazie alla *numerazione univoca* invece sappiamo che un nodo avrà sempre la stessa posizione, **eliminando eventuali ambiguità nella rappresentazione degli elementi nella struttura**.

Dunque, le metodologie ad albero aiutano a risolvere un problema tradizionale di **regressione**: infatti nell'analisi è presente una variabile di risposta numerica, o anche

categorica (qualitativa) **y** e un insieme **X1, X2, Xp** di altre covariate o variabili esplicative o variabili predittive con l'obiettivo di stabilire una relazione tra la variabile di risposta e i predittori allo scopo di predire il valore assunto dalla y in base ai valori assunti dalle variabili esplicative.

Questo quindi è un classico problema di regressione, modello di tipo analitico. Nel caso delle metodologie d albero, invece, non abbiamo un modello analitico ma un metodo **computazionale** che dà luogo a un modello grafico.

I **Vantaggi** delle procedure ad albero si possono riassumere in quattro caratteristiche principali:

- **Metodologie non parametriche** e quindi non si basano su principi distribuzionali. Non bisogna fare nessuna ipotesi per applicare la metodologia ad albero, in quanto si tratta di metodi *data-driven*, guidati dai dati inoltre risultano più utili quando le relazioni sono di tipo non lineare, differentemente utilizzeremmo gli strumenti che già conosciamo (il modello di regressione).
- **Flexible.** In quanto rendono possibile l'utilizzo congiunto di variabili esplicative sia di tipo numerico che categorico, con i predittori presi uno per volta.
- **Powerful.** Ci consentono di analizzare grandi insieme di dati in pochissimo tempo.
- **Simple.** La rappresentazione grafica è di facile comprensione

Alcuni aspetti tecnici della metodologia **CART** sono di primario interesse, infatti:

- **Il CART non richiede che le variabili siano selezionate in anticipo.**
- **I risultati sono invarianti rispetto a trasformazioni monotone delle variabili indipendenti:** ciò comporta che non c'è alcuna necessità in ambito applicativo di sperimentare trasformazioni monotone delle variabili indipendenti (logaritmi, radici quadrate, elevamento a potenza positiva, etc.). Nel CART tali variazioni non modificano i risultati a meno che lo split sia basato su combinazioni lineari di variabili.
- **Il CART è estremamente robusto all'effetto degli outliers.**
- **Può utilizzare combinazioni lineari delle variabili per determinare gli split.**
- **Può mettere in luce dipendenze ed interazioni.**
- **Può processare casi con dati mancanti.**

- **Produce alberi ottimali.**
- **Utilizza strumenti molto sofisticati per stabilire la sua accuratezza.**
- **Può utilizzare la stessa variabile in punti differenti dell'albero.**
- **Può essere utilizzato a supporto dei modelli parametrici convenzionali**.

Tuttavia alcuni **Inconvenienti** sono stati rilevati:

- Il principale inconveniente è che questa procedura si basa su un algoritmo chiamato **Divide and Conquer**. Questo vuol dire che la suddivisione e Splitting procede e continua fino a quando o un nodo è puro (contiene o una sola osservazione o osservazioni tutte omogenee); oppure quando viene soddisfatta una regola di stop (regola che arresta in maniera artificiale la procedura, ad esempio un nodo non viene più suddiviso se contiene meno di 10 osservazioni). Pertanto al di là delle regole di stop che si possono definire a priori (che sono di per se un male, perché a priori le informazioni contenute nei dati difficilmente risultano note), la **conseguenza** è che la struttura tende ad essere molto grande e complessa, perdendo il vantaggio della semplicità.
- Oltre a perdere il vantaggio della semplicità, c'è il problema **dell'overfit** – sovra adattamento cioè più si scende giù nella struttura più gli split riguardano nodi con poche osservazioni. Osservazioni che hanno delle caratteristiche che possono non presentarsi nella generalità dei casi e nella popolazione.

Va considerata la critica più comune rivolta all'uso degli alberi di decisione nei problemi di classificazione. La critica sostiene che negli alberi di decisione, la decisione della variabile split avviene ad ogni nodo dell'albero, in un preciso momento durante l'esecuzione dell'algoritmo, e non è mai più riconsiderata in seguito. Infatti, quando l'algoritmo sceglie, in un determinato nodo, un opportuno split non è assolutamente detto che quello sia (in assoluto) il miglior split per quel nodo; sicuramente è il miglior split per quel nodo in quel preciso istante ed alla luce delle informazioni che l'algoritmo

possiede in quell'istante; il problema è che, alla luce di nuove informazioni, l'algoritmo non andrà a modificare gli split dei nodi precedenti.

Di conseguenza tutti gli split vengono scelti sequenzialmente e ogni split è dipendente dai precedenti. Ciò implica che tutti i futuri split sono dipendenti dal nodo radice dell'albero, a tal punto che una modifica dello split del nodo radice potrebbe portare alla costruzione di un albero completamente differente.

ELENCO DELLE FIGURE

Bibliografia E Sitografia

D'Urso, P. (2016). *METODI STATISTICI PER IL MARKETING - Università degli Studi LUISS* .

Mentuccia, L. (2016, Ottobre 12). *Automotive, sprint dall'integrazione online-concessionari*. Tratto da Cor.Com: http://www.corrierecomunicazioni.it

Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J.(1984). *Classification and regression trees*, Wadsworth International.

Cappelli C. (2002), La Validazione della struttura nelle metodologie ad albero. Tesi di Dottorato di Ricerca, XII ciclo. Università di Napoli Federico II

C Cappelli, F Mola, R Siciliano (2002). A statistical approach to growing a reliable honest tree.. Computational statistics & data analysis 38 (3), pg 285-299.

Buntine, W. (1992). *Learning classification trees, Statistics and Computing*, 2, 63-73.

Efron, B. (1983). *Estimating the error rate of a prediction rule: improvement on cross-validation*, Journal of the American Statistical Association, 78, 316-330.

Hastie, T. Tibshirani, R. and Friedman, J. (2001) *The elements of statistical learning: data mining, inference and prediction*. Springer, New York.

Kass, G. (1980) *An exploratory technique for investigating large quantities of categoriucal data. Applied Statistics*, 29, 119-127.

Morgan, J.N. & Sonquist, J.A. (1964). *Problems in the analysis of survey data and a proposal, Journal of American Statistical Association*, 58, 415-434.

Quinlan, J.R. (1986). *Induction of decision tree*, Machine Learning, 1, 86-106.

Sonquist, J.A. & Morgan, J.N.. (1963). *The detection of interactions effects. Ann Arbor: Institute for Social Research,* University of Michigan.

Paul Gao, Hans-Werner Kaas, Detlev Mohr, and Dominik Wee, *Disruptive trends that will transform the auto industry*, http://www.mckinsey. com/industries/high-tech/our-insights/disruptive-trends-thatwill- transform-the-auto-industry, *McKinsey, 2015*

KPMG, KPMG's Global Automotive Executive Survey 2015, *Who is fit and ready to harvest?*, UK, USA, Europe, 2015

Nawangwulan, I. M., Anantadjaya, S. P. D., Widayatmoko, D. H., and Seancho, W. M., *Consumer Behaviors and Customer Satisfaction: Any Value Created?, International Conference organized by the Society for Interdisciplinary Business Research and Thammasat University*, Bangkok, 2011

Dulli Susi; Furini Sara; Peron Edmondo. *Data Mining*, Springer Verlag, 2009

Rokach, Lior; Maimon, *O. (2008). Data mining with decision trees: theory and applications*. World Scientific Pub Co Inc.

Quinlan, J. R., (1986). Induction of Decision Trees. Machine Learning 1: 81-106, Kluwer Academic Publishers

Friedman, J. H. (1999). Stochastic gradient boosting. Stanford University

Breiman, L. (1996). Bagging Predictors. "Machine Learning, 24": pp. 123-140
Murthy S. (1998). Automatic construction of decision trees from data: A multidisciplinary survey. Data Mining and Knowledge Discovery.

"Management Tools - Customer Relationship Management - Bain & Company". www.bain.com. Retrieved 23 November 2015.

Shaw, Robert (1991). Computer Aided Marketing & Selling. Butterworth Heinemann. ISBN 978-0-7506-1707-9.

David Sims, TMC.net (2007) CRM Adoption 'Biggest Problem' in 83 Percent of Cases.

Jim Lecinski, *Winning the Zero Momento of Thruth, Winning the shift to mobile*, Google, 2015

Netpop Research, Global Auto Shopper Study, Italy, 2015

Nawangwulan, I. M., Anantadjaya, S. P. D., Widayatmoko, D. H., and Seancho, W. M., *Consumer Behaviors and Customer Satisfaction: Any Value Created?,* International Conference organized by the Society for Interdisciplinary Business Research and Thammasat University, Bangkok, 2011

Magnaghi Marco, Email, Social Media, e Web: Creare nuove relazioni con i clienti, Hoepli, 2014

James P. Womack, Daniel T. Jones, and Daniel Roos, *The Machine That Changed the World: The Story of Lean Production, Toyota's Secret Weapon in the Global Car Wars That Is Now Revolutionizing World Industry Paperback use pre formatted date that complies with legal requirement from media matrix*, London, 2007

Prandelli, *Marketing in Rete*, McGraw Hill, Verona, 2006

Berry, L. L, Carbone L.P., and S.H. Haeckel, *Managing the Total Customer Experience*, MIT Sloan Management Review, San Francisco, 2002

Customer relationship management (CRM) in business-to-business (B2B) e-commercenull". Information Management & Computer Security. 11 (1): 39–44. 1 March 2003. ISSN 0968-5227. doi:10.1108/09685220310463722.

"Unlock the Mysteries of Your Customer Relationships". Harvard Business Review. Retrieved 22 November 2015.

What is customer relationship management (CRM) ? - Definition from WhatIs.com". SearchCRM. Retrieved 22 November 2015

Reinartz, Werner; Krafft, Manfred; Hoyer, Wayne D. (August 2004). "*The Customer Relationship Management Process: Its Measurement and Impact on Performance*". Journal of Marketing Research. 41 (3): 293–305

Types of CRM and Examples | CRM Software". www.crmsoftware.com. Retrieved 22 November 2015.

William Pride, O.C. Ferrel: Marketing. McGraw-Hill, 2005

Romano, A.; Schael, T.: Il CRM come leva competitiva nelle TLC. VoiceCom News, n. 2, aprile-giugno 2005: 43-46, Milano, 2005

Schael, T.: Il CRM in Italia. Sistemi & Impresa, n.5, giugno 2005: 45-51, Milano 2005

Schael, T.: Come gestire i clienti della piccola impresa italiana? VoiceComNews, n. 1, gennaio-marzo 2008: 52-57, Milano, 2008

Schael, T.: Customer Experience. VoiceComNews, n. 2, aprile-giugno 2008: 10-14, Milano, 2008, slideshare.net.

Schael, T.: Gestirete – il CRM per la microimpresa. VoiceComNews, n. 4, ottobre-dicembre 2008: 53-56, Milano, 2008, slideshare.net.

Schael, T.: Customer Experience Management. VoiceComNews, n. 1, gennaio-marzo 2009: 12-22, Milano, 2009, slideshare.net.

Giaccari, F. Giaccari, M., CRM Magazine: Cosa è il CRM, CRM magazine, Lecce, 2009

Paglialonga, A. : Cos'è il CRM?, Angelo Paglialonga, Foggia, 2014

Maurizio Duse: Il CRM strategico. Come migliorare la competitività aziendale fidelizzando e centralizzando il cliente. Franco Angeli. Milano. 2011

RIASSUNTO

Questo testo affronta il tema della Classificazione Statistica ovvero il problema della predizione di variabili categoriche con l'utilizzo di altre grandezze note (numeriche o categoriche). Inoltre è riportata un'analisi effettuata durante la mia esperienza di Stage in PSA Retail Roma grazie a cui ho potuto apprendere come il mondo dell'automotive durante gli ultimi anni sia stato fortemente caratterizzato dall'influenza del Web. Dunque ho analizzato in RStudio un dataset riguardante le richieste online effettuate dagli utenti nel primo trimestre del 2017, utilizzando il pacchetto rpart ( ) per implementre la costruzione dell' albero di classificazione.

I CART, cioè Alberi di Regressione e Classificazione, sono una delle metodologie utilizzate per implementare il tema in esame. Teorizzati empiricamente negli anni '50 principalmente in realtà militare e medica, sono stati successivamente descritti per la prima volta nel 1984 da Breiman tramite la stesura e pubblicazione del libro CART (Classification and Regression Trees), considerato ancora oggi in letteratura, un punto di riferimento.

Vista la crescente complessità dei fenomeni e dell'analisi di dati non trattabili adeguatamente con tecniche tradizionali e di dati non standard (incompleti, per i quali il numero di componenti non risulti fissato o risulti molto elevato), le metodologie ad albero riscuotono sempre più interesse in ambito scientifico grazie anche ai progressi informatici che hanno reso automatica la loro applicazione.

Quell'ambito della statistica che si serve di informazioni completamente o parzialmente note per dedurre il valore di altre grandezze di nostro interesse invece non note, sfocia nella teoria della predizione.

Parlando di teoria della predizione ci si riferisce prettamente a una nozione generale, rappresentata dalla Classificazione quando la classe è un valore nominale; Regressione quando la classe è un valore numerico. Questi processi hanno come obiettivo la creazione

di modelli che permettono di studiare e descrivere gli insiemi di dati e attuare previsioni per il futuro.[14]

Un CART non è nient'altro che un predittore del valore di una variabile di risposta (target) in funzione di un insieme di variabili indipendenti (input).

Utilizzando una terminologia più precisa chiameremo:

• variabili di misurazione le grandezze predittive (**gli input**) che possono essere sia quantitative che qualitative;

• variabile di risposta la grandezza predetta (**l'output**).

Il modello è strutturato secondo un diagramma ad albero e seconda della natura della variabile target, possiamo parlare di:

• Alberi di Regressione (Regression Trees, la variabile target è quantitativa)

• Alberi di Classificazione (Classification Trees, la variabile target è qualitativa).

Parleremo di regressione nel caso in cui la variabile risposta assuma valori continui, di classificazione nel caso in cui i valori assumibili dalla variabile risposta siano categorici (si tratta di classi di appartenenza).

Il termine Classificazione può indicare due problematiche piuttosto simili che storicamente hanno assunto la stessa denominazione ma allo stesso tempo vanno specificate distintamente.

3. Il primo significato del termine Classificazione fa riferimento a quell'insieme di metodologie statistiche che aspirano ad assegnare una classe ad un dato di classe sconosciuta sulla base delle informazioni fornite da un campione di dati di classe invece nota (Learning sample).
4. Il secondo significato del termine Classificazione invece sta ad indicare tutte quelle tecniche che hanno come obiettivo l'individuazione di classi o gruppi più o meno evidenti in un insieme di dati non classificati a priori (Cluster Analysis).

Si nota come le due definizioni siano apparentemente simili ma si differenzino in base ai loro scopi infatti:

- la Classificazione (nel senso di assegnazione) cerca di capire in uno Spazio X a quale classe di assegnazione appartenga ogni dato da classificare.

[14] Hastie, T. Tibshirani, R. and Friedman, J. (2001) The elements of statistical learning: data mining, inference and prediction. Springer, New York

- la Cluster Analysis invece cerca in uno Spazio *X* di capire se esistono delle classi di assegnazione.

I modelli strutturati ad albero sono stati usati in campi diversi come la botanica e la medicina, prima che il loro potenziale esplicativo venisse scoperto dagli statistici, specialmente per applicazioni relative a problemi di classificazione.
Gli alberi di classificazione (o di s*egmentazione*) rappresentano una metodologia che ha l'obiettivo di ottenere una segmentazione gerarchica di un insieme di unità statistiche mediante l'individuazione di "regole" (o "percorsi") che sfruttano la relazione esistente tra una classe di appartenenza e le variabili rilevate per ciascuna unità, dunque sono il risultato di un processo di suddivisione ricorsiva di un insieme di unità statistiche in sottogruppi disgiunti, caratterizzati da un grado di omogeneità crescente e di numerosità via via inferiore[15].

Formalmente un albero è costituito da un insieme finito di elementi detti **nodi**; il nodo da cui si diramo i successivi viene detto **radice**.

[15] Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J.(1984). Classification and regression trees, Wadsworth International

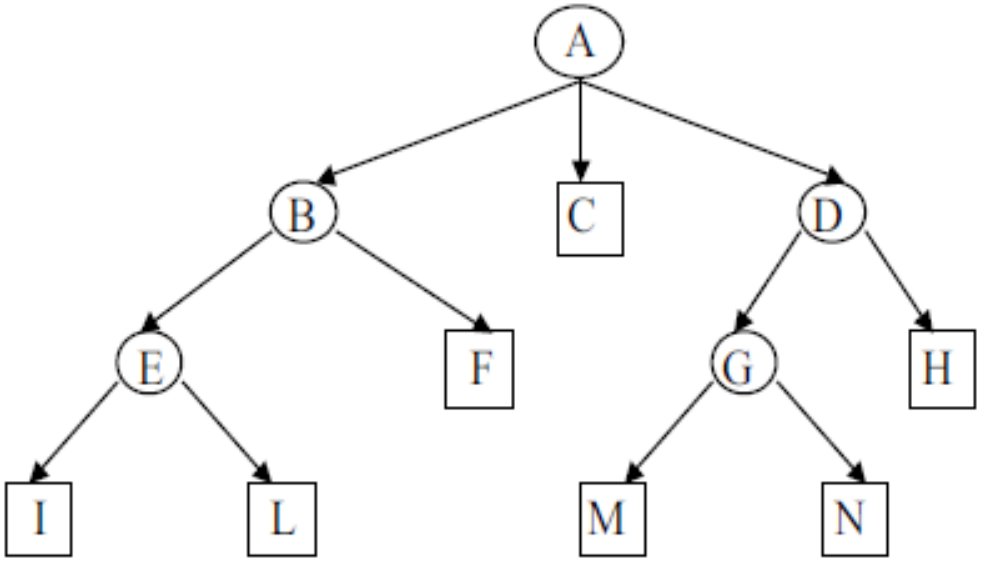

Figura 25- Esempio di generica struttura ad albero- Non parametric regression and classification theroy and an application to web-marketing- Prof. C.Cappelli, 2016

L'albero rappresentato in *Figura 12* presenta 4 livelli, partendo dal nodo radice (A) che costituisce il primo livello, si incontrano due nodi *interni* (B,D) nel secondo livello insieme ad un nodo cosiddetto *terminale* o *foglia* indicato dal riquadro (D). Passando al terzo livello sono riportati ancora una volta due nodi interni (E,G) e due terminali (F,H) ed infine nel quarto livello sono rappresentati i nodi terminali (I, L, M, N) con cui termina la struttura non essendo ulteriormente suddivisi. Si noti come ciascun nodo interno, rappresentato da un cerchietto, sia a sua volta radice di un sottoalbero o *branca*: potando la branca che deriva da un qualunque nodo interno, si ottiene un sottoalbero dell'albero originario, che presenta la stessa radice.

Gli split sono i valori soglia che dividono le unità di un determinato nodo. Il modo più semplice ed efficace per classificare osservazioni in un numero finito di classi sono gli alberi di decisione che vengono realizzati suddividendo più volte le osservazioni in sottoinsiemi omogenei rispetto alla variabile dipendente o risposta.

Questo metodo crea una gerarchia ad albero, dove i sottoinsiemi iniziali sono i nodi(e quelli finali le foglie.

Nello specifico, i nodi sono etichettati con il nome delle variabili, gli archi, cioè i rami dell'albero, con i possibili valori della variabile soprastante, e le foglie dell'albero con le diverse modalità della variabile "classe" che descrivono i gruppi di appartenenza.
Così facendo ogni oggetto viene classificato seguendo un determinato percorso lungo l'albero, dalla radice ad una precisa foglia. I percorsi sono rappresentati dai rami dell'albero che forniscono le regole base per orientarsi nel sistema.
Dalla Figura 11 notiamo come ciascun nodo interno dia luogo ad un numero variabile di suddivisioni; nel caso il numero delle variabili sia pari a 2 e resta costante per ogni nodo interno e per ogni livello, ci troviamo difronte a degli alberi binari, importanti perché le metodologie ad albero tipicamente operano per dicotomie all'interno del dominio fondamentale in esame dando luogo ad alberi di tal tipo, pertanto dette di segmentazione binaria. [16]
Dall'analisi appena fatta è evidente come il ruolo degli alberi di decisione sia importante tanto quanto utile per il supporto di determinate decisioni e la rappresentazione di diversi tipi di informazione; lampante è l'esempio dell'albero genealogico anche se moltissime altre informazioni sono rappresentate con una struttura ad albero e tante altre si presterebbero alla medesima rappresentazione Questo tipo di strumento ha il vantaggio di esprimere graficamente i concetti di progressività e di inclusione e di conseguenza di gerarchizzazione, infatti consente di rappresentare legami gerarchici tra i dati.
Le metodologie ad albero sono utili ai fini di predire la variabile *Y* detta *variabile di risposta* o *criterio* sulla base dei valori assunti da un insieme di predittori *X*1*;X*2*; : : : ;XK*.
Nel caso in cui la variabile di risposta sia *categorica*, riferendoci alle metodologie ad albero possiamo parlare di metodo di Classificazione, diversamente in caso di variabile criterio di tipo *numerico* si parla di *Regressione ad albero*.
In entrambi i casi il fine ultimo non è altro che l'assegnazione alla variabile *Y* di un valore che rappresenta un'etichetta corrispondente ad una classe.
Tecnicamente, si prende in esame la probabilità:

$$f\,(Y = y|\, X_1, X_2, \ \dots\ , X_K); \qquad (2.2.1)$$

Nel caso in cui *Y* sia numerica si considera il valore atteso:

$$E(Y \mid X_1, X_2, \ \dots\ , X_K). \qquad (2.2.2)$$

[16] Buntine, W. (1992). Learning classification trees, Statistics and Computing, 2, 63-73.

In ambito statistico se esaminassimo la variabile Y dicotomica o politomica, ricorreremmo alla regressione logistica [17], e nel caso in cui i predittori fossero tutti di tipo numerico utilizzeremmo l'analisi Discriminante lineare di Fisher o le sue estensioni e varianti quali ad esempio l'analisi discriminante quadratica, che considera interazioni di ordine superiore tra i predittori, infine se prendessimo in esame una variabile di risposta numerica la soluzione standard verrebbe fornita dal modello di regressione lineare.

Tali metodi tradizionali si fondano su ipotesi fortemente restrittive riguardanti la forma distribuzionale dei dati o il tipo di legame esistente tra la variabile di risposta ed i predittori. Di conseguenza si otterrebbero risultati poco affidabili con un'interpretazione dei dati e del modello in esame decisamente problematica.

In contrasto con tali metodi classici, le metodologie ad albero presentano dei notevoli vantaggi:

- Essendo tecniche non parametriche non necessitano di un modello specifico;
- I predittori da utilizzare possono risultare di diversa natura;
- La rappresentazione grafica che ne scaturisce risulta di facile interpretazione in quanto consente di visualizzare le relazioni esistenti tra variabile di risposta e predittori in maniera immediata.

Si potrebbe dire che tali metodologie rispondono ad un problema classico della statistica senza presentare molti degli inconvenienti dei metodi classici impiegati al medesimo scopo. [18]

Come anticipato nel paragrafo precedente il sistema di rappresentazione ad albero di classificazione, può avere svariati utilizzi perché la loro struttura ha il vantaggio di poter essere memorizzata in modo uniforme, di aggiungere alla classificazione eventuali nuove informazioni, e sono applicabili per risolvere problemi di diversa natura.

Tra gli svantaggi invece troviamo la difficoltà nel trovare un albero di dimensione ottimale nonostante esistano metodi "top-down" che partono dalla radice ripartendo solo successivamente lo spazio delle variabili.

Per costruire una struttura ad albero efficace occorre seguire 3 step:

[17] In tal caso la (2.2.2) coincide con la (2.2.1) con y= 1.

[18] Cappelli C. (2002), La Validazione della struttura nelle metodologie ad albero. Tesi di Dottorato di Ricerca, XII ciclo. Università di Napoli Federico II

4. Selezionare una regola di splitting per ogni nodo, ciò significa determinare le variabili, insieme al rispettivo valore soglia, che saranno usate per partizionare il data set ad ogni nodo.
5. Determinare quali nodi sono da intendersi terminali, quindi per ogni nodo bisogna decidere quando continuare con gli splits, quando fermarsi e considerare il nodo come terminale e di conseguenza assegnargli un'etichetta. Infatti senza un'adeguata regola, si corre il rischio di costruire alberi troppo grandi con una piccola capacità di generalizzazione, oppure alberi troppo piccoli che invece approssimano male i dati.

6. Assegnare le etichette ad ogni nodo terminale, ad esempio minimizzando il valore atteso di errata classificazione. A partire da un qualunque tipo di problema è quasi sempre possibile costruire l'albero di decisione corrispondente.

Lo scopo finale è l'individuazione di relazioni esistenti tra la variabile di risposta ed i predittori, cosicché sia necessario individuare un insieme di regole di classificazione/predizione, nella forma di un albero binario.Viene utilizzato un insieme di apprendimento, $C = \{(y_i, x_i, i = 1, \dots , n)\}$ ovvero un insieme di osservazioni su cui sono state rilevate i valori assunti dalla variabile criterio e dai predittori ; l'insieme di casi si supponga essere estratto da una variabile casuale multivariata (X; Y) dove X è il vettore dei K predittori ed Y è la variabile di risposta.
Questo insieme rappresenta l'esperienza passata e costituisce ciò che si conosce, il cosiddetto dominio fondamentale, impiegato appunto per creare l'albero, con la duplice valenza, di esplorare i dati dell'insieme di apprendimento e di indurre il valore di risposta da associare a nuove osservazioni che scivolando nella struttura raggiungono un nodo terminale.
Si consideri un esempio di output di dati raccolti negli USA, in cui la variabile di risposta (y) è il reddito annuale, suddiviso in due classi: >50.000$ e <50.000$.
Le variabili esplicative sono: l'età (anni passati a scuola), relazione familiare, e il capital gain.

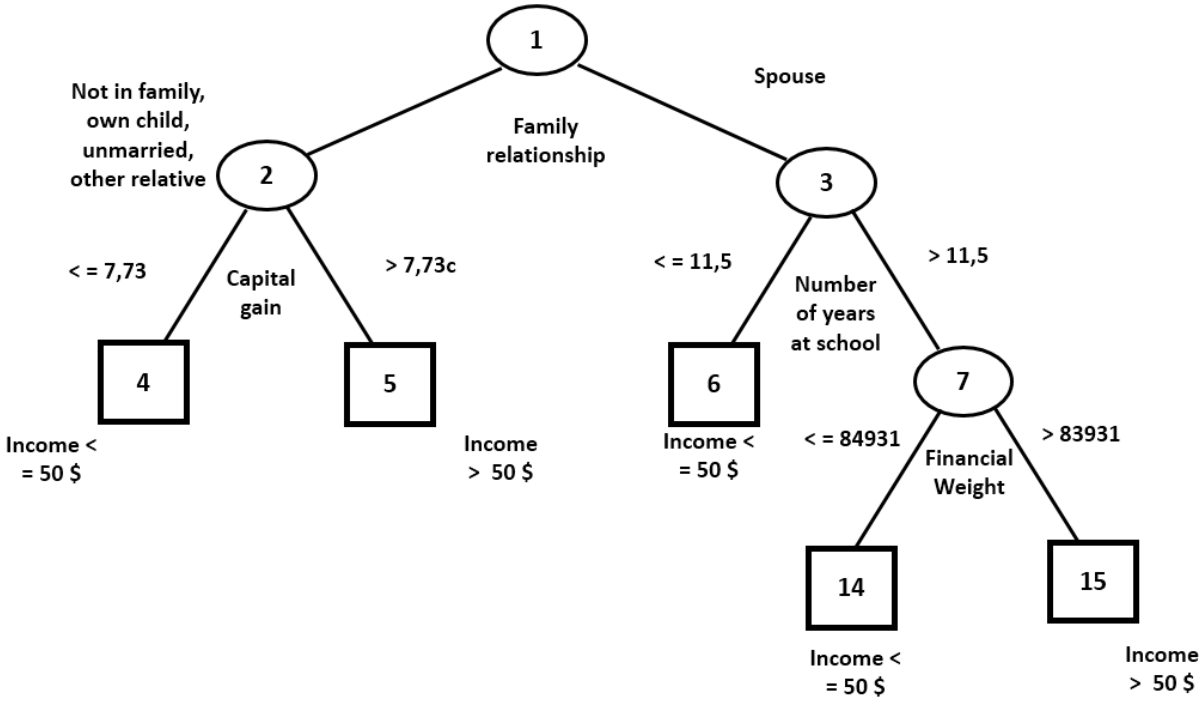

Figura 26- Albero binario Ipotetico- Non parametric regression and classification theroy and an application to web-marketing- Prof. C.Cappelli, 2016

All'inizio della procedura nel nodo 1 che è il nodo radice da cui tutta la struttura nasce, sono presenti tutti i rispondenti, dunque tutte le osservazioni.

Si nota come la variabile (nell'ambito delle variabili esplicative) che maggiormente suddivide queste osservazioni in due sottogruppi più o meno omogenei rispetto al reddito annuale è quella delle Relazioni Familiari. Nel **nodo 2** (nodo figlio a sinistra) vanno a finire tutti i rispondenti che non sono in famiglia, *unmarried*, o vivono con altri familiari; nel **nodo 3**, sono presenti i rispondenti in una condizione *spouse*, coniugale. Quindi in base alla variabile *relazione familare* i rispondenti sono stati divisi in due sottogruppi che sono più omogenei rispetto all' unico insieme iniziale e sono omogenei in termini di reddito annuale che percepiscono.

Nel **nodo 3** la variabile che risulta maggiormente esplicativa, che suddivide i rispondenti *spouse* in sottogruppi ancora più omogenei, è il numero di anni a scuola.

In particolre, nel **nodo 6** abbiamo tutti i rispondenti in condizione *spouse* che però non hanno un grado di istruzione elevata, e queste persone percepiscono un reddito modale < 50.000$. Quelli che invece hanno raggiunto più della soglia di 11,5 in termini di

numero di anni scolastici finiscono nel **nodo 7** in cui la variabile che li discrimina in sottogruppi ancora più omogenei è *financial weight*

Questa struttura grafica quindi sulla base delle covariate ci dà una serie di possibili regole di predizione, e queste regole sono 5, quanti sono i nodi terminali. Perché corrispondono a diversi percorsi che dal nodo radice conducono a un gruppo omogenei rispetto alla variabile di risposta.

Si noti come i nodi siano designati secondo una numerazione binomia che, a partire dal nodo radice cui è assegnato il numero uno, assegna ai discendenti di un nodo interno un numero identificativo pari all'identificativo del nodo padre moltiplicato due per il discendente di sinistra e con l'aggiunta di una unità per il discendente di destra. Dunque, questo sistema di designazione risulta estremamente semplice ma efficace, infatti consente di risalire in maniera univoca alla posizione di un nodo nella struttura a partire dal numero ad esso assegnato.

Il punto cruciale è la scelta di queste suddivisioni, delle variabili che permettono di suddividere le osservazioni in due sottogruppi.

Differentemente dal modello di regressione ($y= ß_0 + ß_1x1 + ß_2x2 ...+ Ɛ$), nel modello considerato le covariate non compaiono come sono, ma **compaiono sotto forma di variabili dicotomiche** (family relation: Spose or not spose?...).

Infatti tutta la struttura ruota attorno a queste suddivisioni binarie, più genericamente rappresentano domande binarie del tipo: Xj , che è un generico predittore, appartiene ad A, sì o no? -Risposta dicotomica- Xj è quindi un fattore, una variabile qualitativa, come nel nostro esempio *family relationship*, che non assume un fattore numerico, le cui modalità sono degli attributi e di conseguenza l'insieme A non è altro che un insieme delle modalità del predittore.

Se invece Xj fosse una variabile numerica, il sottoinsieme A sarebbe definito su un intervallo di valori che la variabile numerica può assumere (età a scuola come nel nostro esempio).

Il numero di variabili di suddivisione che si possono generare partendo dalle variabili originarie e quindi dalle coviariate, è un numero che dipende dalla natura, dalla scala di

misura delle stesse, quindi è importante sapere se la variabile iniziale che viene presa in esame è in origine dicotomica, qualitativa ordinale, qualitativa nominale – non sussiste una relazione di ordine – oppure numerica. Quindi ogni covariata può generare un numero di split che dipende dalla sua scala di misura.

Questo specchietto riassume il numero di Split che un predittore può genare:

- Predittore **dicotomico:** (Maschio o Femmina). C'è **1** sola possibile suddivisione (o maschio o femmina).
- Predittore **numerico:** esistono *n* valori distinti che possono generare **n-1** suddivisioni binarie
- Predittore **qualitativo**, ma **c'è** una linearità d'**ordine** nelle modalità. Ad esempio i voti a scuola: sufficiente, buono, distinto, ottimo. Se si volesse splittare in due gruppi sulla base di questi livelli, si potrebbe dividere in un gruppo *sufficiente* e in un altro gruppo *buono-distinto-ottimo*, oppure *sufficiente-buono* in un unico gruppo e distinto-ottimo nell'altro, o ancora da un lato s-b-d e dall'altro ottimo. Quindi si avranno 3 possibili suddivisioni binarie a partire dai 4 livelli del predittore preso in considerazione; riassumendo: le modalità sarebbero **m** =4 e le suddivisioni sarebbero **m-1** =3
- Il caso più complicato dal punto di vista computazionale è quello dei fattori, delle variabili **qualitative** che **non hanno una relazione d'ordine** tra le modalità (*titolo di studio superiore: liceo classico, scientifico, linguistico*…etc sono attributi tra cui non esiste una relazione d'ordine). Si potrebbero raggruppare sotto forma di combinazioni lineari in molti modi; considerate m modalità della variabile, è possibile generare ben $\mathbf{2^{(m-1)} - 1}$ variabili di suddivisione, split (dunque con un grande peso computazionale).

Si pone dunque un primo problema nella costruzione dell'albero, ossia scegliere quale predittore o partizione binaria debba essere utilizzata per segmentare ciascun nodo; per quale principio i nodi terminali non vengano segmentati e come viene associato ciascun nodo terminale ad una classe di risposta.

Questi interrogativi delineano i punti cardine delle metodologie ad albero così classificati:

5. Un insieme di domande binarie risultanti dalla dicotomizzazione dei predittori;

6. Una regola, cosiddetto *criterio di split* che consenta ad ogni nodo interno di scegliere la domanda binaria in base alla quale suddividere il nodo;
7. Un criterio che consenta di dichiarare un nodo terminale;
8. Una regola di assegnazione di un valore di risposta a ciascun nodo terminale.

Al fine di dividere il campione in sottogruppi e decidere quale valore soglia utilizzare sulle variabili, è necessaria una regola di splitting per decidere quale variabile, o combinazione di variabili, deve essere usata in un certo nodo.

La procedura di segmentazione binaria consiste in una partizione binaria ricorsiva di casi N in due sottogruppi disgiunti. Lo scopo della procedura è quello di definire una regola di classificazione / predizione sulla base di un set di apprendimento (chiamato anche training set), per i quali sono stati registrati valori di una variabile di risposta Y e di una serie di variabili esplicative XI, X2, ..., XK (sia numeriche che categoriche).

I dati sono partizionati scegliendo ad ogni passo una variabile e un punto di taglio lungo di essa in base a una bontà di misura divisa che consente di selezionare quella variabile (più precisamente una variabile di divisione) che genera i sottogruppi più omogenei rispetto alla variabile di risposta.

La procedura di partizionamento ricorsiva segue un algoritmo "divide and conquer", nel senso che in linea di principio l'algoritmo continua i nodi di partizionamento finché tutte le foglie contengono un singolo caso oi casi che appartengono alla stessa classe o che presentano lo stesso valore di risposta.

Di conseguenza, un ulteriore passo, la semplificazione dell'albero, è di solito effettuata per evitare l’overfitting e migliorare la comprensione dell'albero rimuovendo retrospettivamente alcuni dei rami (la cosiddetta potatura)

Una volta che la struttura ad albero è stata potata, ad ogni nodo terminale viene assegnata un'etichetta di classe o un valore di risposta

Riepilogando, i metodi basati sull'albero coinvolgono i seguenti passaggi:

1) crescita dell'albero

2) potatura dell’albero

3) assegnazione delle classi / valori di risposta ai nodi terminali.

Il punto è quello di scegliere tra le variabili di split la migliore, che garantisca la suddivisione delle osservazioni presenti nel nodo in sottogruppi il più possibili omogenei al loro interno ed eterogenei tra loro.

Occorre effettuare una sorta di **valutazione della bontà** degli *split*, dunque è necessaria una misura della loro qualità in termini di riduzione della eterogeneità, della variabilità della variabile di risposta in modo da poter scegliere lo *split* migliore (che suddivide le osservazioni in due sottogruppi che sono in assoluti i migliori, rispetto ad ogni altro possibile *split*, e il più omogenei all'interno); in letteratura si parla a tal proposito di ***Goodness of Split Measure.***[19]

Questo tipo di misure sono state proposte nell'ambito dell'intelligenza artificiale, e non nell'ambito della statistica, la prima in assoluto è stata proposta nel campo del *machine learning*, considerando una variabile di risposta di tipo numerico. Qualche anno dopo, uno studioso di intelligenza artificiale ha proposto una variabile di questo metodo per il caso in cui la variabile fosse stata non numerica e quantitativa, ma qualitativa e che denotava quindi l'appartenenza a dei gruppi.

In seguito queste procedure sono state esaminate anche in campo statistico e tal proposito gli autori del CART hanno sviluppato un framework metodologico, introducendo il concetto generico di ***Impurity*** (impurità) che ingloba in sé sia quello di variabilità associato al caso di una variabile numerica, e sia quello di eterogeneità o mutabilità associato al caso di una variabile di risposta categorico.

Considerando la fase di generazione dell'albero, si parte dalla totalità delle osservazioni appartenenti al training set e si procede con una divisione binaria in classi. Si deve però stabilire preliminarmente il metodo in base al quale effettuare gli *splits*, basato sulla definizione di **impurità.**
L'impurità è una funzione della frazione delle osservazioni classificate in ciascuna classe. Ad esempio, se *N* è il numero delle osservazioni contenute nel training set, una prima divisione binaria delle unità porterebbe alla formazione di due classi contrassegnate con le etichette "1" e "2", ognuna delle quali possiede una determinata frazione della totalità delle osservazioni. Indicate con F_1 e F_2 tali frazioni, la funzione di impurità I(F_1 , F_2)

[19] Quinlan, J.R. (1986). Induction of decision tree, Machine Learning, 1, 86-106

può essere pensata come una funzione, avente valori in [0,1], che fornisce una misura di quanto le osservazioni siano correttamente distribuite nelle classi.

Essa è definita in modo tale che $I = 1$ se le osservazioni sono concentrate in una sola classe e $I = 0$ se esse sono divise in parti perfettamente uguali fra le due classi. Poiché l'obiettivo dell'algoritmo è quello di posizionare le osservazioni in classi prestabilite, esso cercherà effettuare questa classificazione nel modo più corretto possibile e non considererà quegli attributi o quei valori degli attributi che non portano ad una corretta classificazione delle osservazioni. Formalmente, l'algoritmo nelle varie fasi sceglierà degli attributi splitter tra quelli presenti nel training set e proverà diversi valori in modo che in ogni nodo venga minimizzata la funzione di impurità. Minimizzare tale funzione coincide col trovare nelle varie fasi gli attributi e il rispettivo valore soglia, che operano la classificazione più corretta possibile e che quindi dovrebbero dare in quella fase un'informazione maggiore.

Il procedimento è di tipo *iterativo* e si arresterà quando non sarà più possibile, manipolando la scelta degli attributi e/o il loro valore soglia, diminuire la funzione di impurità.

La **misura di impurità** per il generico nodo t è così definita:

$$\mathrm{I}(t) = \phi(p(1|t), \cdots, p(J|t))$$

Essa è quindi una funzione $\phi(.)$ non negativa di $\{p(j|t)\}$ tale che:

4. $\phi(.)$ è massima solo quando $p(j|t) = \frac{1}{J}$ per ogni j;
5. $\phi(.)$ è minima e pari a zero quando $\phi(1,0,\ldots,0) = \phi(0,1,0,\ldots,0) = \ldots = \phi(0,\ldots,0,1) = 0$;
6. $\phi(.)$è una funzione simmetrica di $p(1|t),\ldots, p(J|t)$;

Quindi l'impurità di un nodo è massima quando tutte le classi della variabile dipendente sono presenti nella stessa proporzione, mentre è minima quando il nodo contiene casi appartenenti ad un'unica classe.

I due metodi comunemente usati per commisurare l'impurità sono l'entropia e l'indice di Gini.

Lo studio da me affrontato propone un'ottica certamente non deterministica, ma offre un'esposizione del cambiamento nel processo d'acquisto nel settore, supportato da

un’analisi del lead management all’interno di un dealer PSA, utilizzando la metodologia di classificazione ad albero.
Entro il 2020, la digitalizzazione in crescita e gli avanzamenti tecnologici avranno aumentato gli investimenti del settore automobilistico a 82 miliardi di dollari. L'industria automobilistica ha imparato rapidamente che bisogna soddisfare le richieste dei consumatori con un'esperienza digitalmente migliorata; i nuovi modelli di business potrebbero espandere il reddito automobilistico di circa il 30%, raggiungendo una soglia di $1,5 trilioni, tuttavia complessivamente le vendite di automobili hanno mostrato un trend di crescita dal 2005 in poi, esclusi il biennio 2008-2009 (che ha risentito della crisi economica), con un incremento delle vendite dal 2007 al 2016 quasi del 29,2%.

Si evince dunque come il settore Automotive abbia subito molte variazioni nell’ultimo decennio, tanto che, dovendo rispondere alle attese del cliente, ogni sistema è stato integrato con una soluzione digitale.
Con le tecnologie digitali**,** i consumatori stanno cambiando il processo d’acquisto. Quando si recano presso le concessionarie automobilistiche, i clienti cercano informazioni specifiche, quelle che non trovano su Internet, e più che una presentazione dell’auto, hanno bisogno di consigli da parte di professionisti che abbiano una forte preparazione sul prodotto e sulla gestione dell’esperienza con il cliente**.**
La Costumer Journey non inizia sulla soglia degli showrooom, ma sui touch point virtuali, quindi la prima sfida da affrontare consiste in una rivisitazione del processo di vendita. Di fondamentale importanza rappresenta il raggiungimento della consapevolezza che, dietro all’immaterialità di un contatto web, c’è sempre una persona e, quindi, un potenziale cliente. Sarebbe inammissibile lasciare al caso il primo approccio del consumatore sul digital, lasciandolo slegato dal resto del processo di vendita, quindi è necessario trovare un punto di collegamento tra la comunicazione sui canali tradizionali e quella sul web.

Si deduce quanto ci sia margine di miglioramento del settore nell’interazione con i clienti, utilizzando nuovi strumenti di CRM i quali non sono altro che applicativi in grado di raccogliere e strutturare dati e tecnologie per offrire esperienze di realtà aumentata e virtuale.

Durante il percorso di stage intrapreso in PSA Retail Italia, esaminando i dati del sito E-Dealer del marchio Peugeot, è risultato che il 70% degli utenti visita solitamente il sito naturalmente, l'8% da campagne e banner, 10% direct tramite Query Code che rimanda al sito E-Dealer, il 12% tramite il Referral; di conseguenza la decisione di non abbandonare alcun canale tradizionale aggiungendo alcune attività di marketing esperienziale con una serie di eventi sul territorio, cercando di ottenere la giusta combinazione tra quello che propone la casa madre a livello nazionale, e quello che comunicano i dealer a livello locale. L'obiettivo di tutta la comunicazione è diventato, quindi, quello di orientare il processo d'acquisto della nuova auto attraverso i presidi sul web, in modo da soddisfare le aspettative del cliente, in termini di risparmio di tempo facilitando la decisione d'acquisto.

Il sito web ha, pertanto, un ruolo ben preciso, infatti rappresenta una vera e trasparente guida all'acquisto. E' stato concepito una sorta di salone virtuale con una strategia in grado di riprodurre esattamente quella costruita per i clienti che visitano gli showroom dei dealer, motivo per cui oggi il cliente può sul sito configurare la vettura d'interesse e ottenere il prezzo finale della stessa.

I sistemi CRM sono utili ad alimentare i database sui clienti attraverso diversi canali o punti di contatto tra il cliente e l'azienda, includendo vari siti web della società e non, telefono, chat, direct mail, materiali di marketing e social media. I sistemi CRM possono anche dare informazioni dettagliate personali inerenti alla cronologia degli acquisti, le preferenze di acquisto e i vari contatti che ci sono stati tra l'azienda e il cliente.

Grazie a questi nuovi mezzi di comunicazione, strumenti fondamentali per la relazione con la clientela, si possono abbattere le barriere spazio temporali, riuscendo a creare una comunicazione simmetrica, interattiva e quasi istantanea.

La peculiarità dei software è quella di convogliare tutte le informazioni in un database CRM unico, in modo che il digital team possa accedervi facilmente. Le altre funzioni principali di questo software comprendono la registrazione di varie interazioni con i clienti (tramite e-mail, telefonate, social media o altri canali, a seconda delle capacità del sistema), automatizzando vari processi di workflow, quali attività, calendari e allarmi, e dando ai manager la possibilità di monitorare prestazioni e produttività in base alle informazioni registrate nel sistema.

Il sistema è fondamentale per snellire il lavoro del digital team, in quanto automatizza molti processi che richiederebbero del tempo, quali l'invio con cadenza periodica di mail e sms, dunque molte procedure che richiederebbero l'ausilio di un dipendente sono ridotte a pochi click.
L'avvento dei social media e la proliferazione dei dispositivi mobili ha portato i fornitori di CRM di aggiornare la propria offerta per includere le nuove caratteristiche che si rivolgono ai clienti che utilizzano queste tecnologie.
L'analisi è effettuata per mezzo del software RStudio considerando il data set precedentemente esposto, quindi la variabile di risposta (y) è rappresentata da Contratto, che indica i Contratti effettuati nel primo trimestre del 2017, in seguito ad una richiesta online effettuata sul sito Web o Facebook. Le variabili esplicative sono: Città (Roma o Fuori Roma), Fonte (Sito web o Facebook e relativa compilazione del format), Preventivo (dicotomica→ sì/no), Tempo (tempo in giorni intercorsi tra la richiesta online e quindi il primo contatto e l'appuntamento in filiale).

Il primo step, dopo aver trasformato il file Excel in formato csv viene caricato in RStudio, caricando il pacchetto 'rpart' viene costruito l'albero di classificazione. Nella costruzione dell'albero sono state usate le variabili 'preventivo', 'tempo' (il periodo di tempo intercorso tra contatto e appuntamento) e 'fonte'; non è stata usata la variabile 'città'.

Abbiamo 8 nodi terminali, identificati dall'asterisco; in ognuno di essi la ripartizione della relativa modalità di contratto è espressa come numerosità e come probabilità tra parentesi

Si nota nel primo nodo (nodo radice), in corrispondenza di preventivo=no, abbiamo 247 soggetti per cui contratto=no e 0 soggetti per cui contratto=sì, una classificazione quindi perfetta a cui corrisponde, nella parentesi, 1 per il no, cioè 100% e 0 per il sì, 0%.

Da questa prima analisi risulta evidente come la variabile, nell'ambito delle variabili esplicative, che maggiormente suddivide le osservazioni in due sottogruppi omogenei sia "Preventivo".

Infatti nella costruzione dell'albero la variabile "Preventivo" ha un'importanza dell'87%, la variabile "tempo" del 10% e la variabile "fonte" del 3%.

Il misclassification rate (o prediction error) può essere calcolato come = Root node error * rel error * 100%

Nel nostro caso misclassification rate= 0.1349*0.76087*100=10.26%, presenta un valore molto basso.

Tramite la funzione `predict` possiamo confrontare la classificazione reale con quella stimata dall'albero; nella tabella leggiamo per riga la classificazione reale e per colonna quella inferita dalle funzioni lineari utilizzate.

Sulla diagonale principale si hanno i casi in cui le classificazioni coincidono. Dunque abbiamo 24+11 casi di classificazione errata corrispondenti ad un error rate di 35/341=10,26%, che coincide con il misclassification rate calcolato in precedenza.

A questo punto per avere una migliore rappresentazione grafica dell'albero costruito utilizziamo il pacchetto **rpart.plot.** Sono presenti 8 nodi terminali, con un errore di classificazione relativo alla modalità di contratto evidenziata e alla percentuale di soggetti rispetto al totale del dataset (ad es. nel secondo nodo terminale c'è il 4% di soggetti totali e classificandoli come contratto=no il tasso di errore è pari a 0,21, cioè al 21%).

Gli alberi di classificazione rappresentano una metodologia che ha l'obiettivo di ottenere una segmentazione gerarchica di un insieme di unità statistiche mediante l'individuazione di "percorsi" che sfruttano la relazione esistente tra una classe di appartenenza e le variabili rilevate per ciascuna unità.

Si parla di albero **binario**, poiché ogni nodo viene diviso in due parti, mediante lo Split.

Tali alberi sono numerati con un sistema di numerazione di tipo binario, utilizzato perché si tratta di ***numerazione univoca***, che ci consente di risalire alla posizione del nodo in una struttura ad albero, qualunque sia la forma della struttura

Se si adottasse una numerazione di tipo *consequenziale* si avrebbe una situazione in cui la numerazione del nodo dipenderebbe dalla posizione del nodo in quella particolare struttura, quindi i nodi sarebbero numerati in modo successivo. Di conseguenza, un nodo non sarebbe univocamente identificato nella struttura e dipenderebbe dalla forma del grafo.

Grazie alla *numerazione univoca* invece sappiamo che un nodo avrà sempre la stessa posizione, **eliminando eventuali ambiguità nella rappresentazione degli elementi nella struttura**.

Dunque, le metodologie ad albero aiutano a risolvere un problema tradizionale di **regressione**: infatti nell'analisi è presente una variabile di risposta numerica, o anche categorica (qualitativa) **y** e un insieme **X1, X2, Xp** di altre covariate o variabili esplicative o variabili predittive con l'obiettivo di stabilire una relazione tra la variabile di risposta e i predittori allo scopo di predire il valore assunto dalla y in base ai valori assunti dalle variabili esplicative.

Questo quindi è un classico problema di regressione, modello di tipo analitico. Nel caso delle metodologie d albero, invece, non abbiamo un modello analitico ma un metodo **computazionale** che dà luogo a un modello grafico.

I **Vantaggi** delle procedure ad albero si possono riassumere in quattro caratteristiche principali:

- **Metodologie non parametriche** e quindi non si basano su principi distribuzionali. Non bisogna fare nessuna ipotesi per applicare la metodologia ad albero, in quanto si tratta di metodi *data-driven*, guidati dai dati inoltre risultano più utili quando le relazioni sono di tipo non lineare, differentemente utilizzeremmo gli strumenti che già conosciamo (il modello di regressione).
- **Flexible.** In quanto rendono possibile l'utilizzo congiunto di variabili esplicative sia di tipo numerico che categorico, con i predittori presi uno per volta.
- **Powerful.** Ci consentono di analizzare grandi insieme di dati in pochissimo tempo.
- **Simple.** La rappresentazione grafica è di facile comprensione

Alcuni aspetti tecnici della metodologia **CART** sono di primario interesse, infatti:

- **Il CART non richiede che le variabili siano selezionate in anticipo.**
- **I risultati sono invarianti rispetto a trasformazioni monotone delle variabili indipendenti:** ciò comporta che non c'è alcuna necessità in ambito applicativo di sperimentare trasformazioni monotone delle variabili indipendenti (logaritmi, radici quadrate, elevamento a potenza positiva, etc.). Nel CART tali variazioni non modificano i risultati a meno che lo split sia basato su combinazioni lineari di variabili.
- **Il CART è estremamente robusto all'effetto degli outliers.**
- **Può utilizzare combinazioni lineari delle variabili per determinare gli split.**
- **Può mettere in luce dipendenze ed interazioni.**

- **Può processare casi con dati mancanti.**

- **Produce alberi ottimali.**
- **Utilizza strumenti molto sofisticati per stabilire la sua accuratezza.**
- **Può utilizzare la stessa variabile in punti differenti dell'albero.**
- **Può essere utilizzato a supporto dei modelli parametrici convenzionali**.

Tuttavia alcuni **Inconvenienti** sono stati rilevati:

- Il principale inconveniente è che questa procedura si basa su un algoritmo chiamato **Divide and Conquer**. Questo vuol dire che la suddivisione e Splitting procede e continua fino a quando o un nodo è puro (contiene o una sola osservazione o osservazioni tutte omogenee); oppure quando viene soddisfatta una regola di stop (regola che arresta in maniera artificiale la procedura, ad esempio un nodo non viene più suddiviso se contiene meno di 10 osservazioni). Pertanto al di là delle regole di stop che si possono definire a priori (che sono di per se un male, perché a priori le informazioni contenute nei dati difficilmente risultano note), la **conseguenza** è che la struttura tende ad essere molto grande e complessa, perdendo il vantaggio della semplicità.
- Oltre a perdere il vantaggio della semplicità, c'è il problema **dell'overfit** – sovra adattamento cioè più si scende giù nella struttura più gli split riguardano nodi con poche osservazioni. Osservazioni che hanno delle caratteristiche che possono non presentarsi nella generalità dei casi e nella popolazione.

Va considerata la critica più comune rivolta all'uso degli alberi di decisione nei problemi di classificazione. La critica sostiene che negli alberi di decisione, la decisione della variabile split avviene ad ogni nodo dell'albero, in un preciso momento durante l'esecuzione dell'algoritmo, e non è mai più riconsiderata in seguito. Infatti, quando l'algoritmo sceglie, in un determinato nodo, un opportuno split non è assolutamente detto che quello sia (in assoluto) il miglior split per quel nodo; sicuramente è il miglior split per quel nodo in quel preciso istante ed alla luce delle informazioni che l'algoritmo

possiede in quell'istante; il problema è che, alla luce di nuove informazioni, l'algoritmo non andrà a modificare gli split dei nodi precedenti.

Di conseguenza tutti gli split vengono scelti sequenzialmente e ogni split è dipendente dai precedenti. Ciò implica che tutti i futuri split sono dipendenti dal nodo radice dell'albero, a tal punto che una modifica dello split del nodo radice potrebbe portare alla costruzione di un albero completamente differente.