Analisi statistiche esplorative su dati dell'accademia alimentare 'Bioimis'

***

Prof. F. De Antoni- Prof. L. Nieddu***
(*) già Università di "Tor Vergata" - Roma, (**) Università degli studi Internazionali- Roma

Roma - Novembre 2018

Indice

Premessa

Con il presente documento si intende restituire all'Accademia Alimentare 'Bioimis' le analisi statistiche effettuate sul set di dati, estratti, dal DB del committente, secondo le indicazioni, dei redattori della presente relazione tecnica. Prima di effettuare delle analisi statistiche si è proceduto alla verifica della qualità del campione al fine di predisporre un insieme di dati elaborabili che possano dare dei risultati con valenza statistica. Nelle fasi preliminari di verifica della qualità dei dati, c'è stata una stretta collaborazione con 'Bioimis' al fine di correggere le eventuali anomalie presenti, individuare le cause di tali anomalie e comprendere i criteri di immissione dati nel DB. Obiettivo della presente relazione si sintetizza nella valutazione dell'efficacia del programma alimentare "Bioimis" utilizzando, inizialmente tecniche statistiche per la valutazione della qualità dei dati e successivamente metodologie rivolte alla individuazione di gruppi di clienti che hanno seguito il programma alimentare in oggetto, sia sotto l'aspetto statico che dinamico. In tutta la fase di analisi dei dati e nei successivi step vi è stata una stretta collaborazione con Bioimis nella persona del dott. Teta a quale sono state inviate tutte le analisi parziali e le spiegazioni in risposta ai quesiti posti dallo stesso. Ciò al fine di redigere una relazione che sia comprensibile e condivisa dal committente. Inoltre si sono avuti nel corso del 2017 incontri nella sede di Cittadella per pianificare e identificare le strategie di analisi e nell'aprile del 2018, nell'ambito dell'incontro annuale Biomis, si è proceduto alla presentazione sintetica dell'intero lavoro effettuato. Durante la presentazione si è convenuto di ripetere alcune analisi sui dati nuovi e completi che al momento della stesura della presente relazione non sono ancora stati inviati. I risultati, delle elaborazioni effettuate sul set di dati, potranno essere utilizzati anche per scopi promozionali della tipologia di approccio dietologico promossa dall'Accademia Alimentare 'Bioimis''. Le analisi sono state effettuate sul file *'anagrafica'* dopo le necessarie modifiche al fine di rendere il data base analizzabile con tecniche statistiche. Pertanto sono state necessarie correzioni dei dati ed esclusioni per mancanza della informazione, ciò ha comportato una riduzione del numero dei casi da sottoporre ad analisi.

1 - I dati

La base dati, sulla quale effettuare le analisi, è stata fornita da 'Bioimis' che ha provveduto ad estrarre, dal proprio DB tre files, in presenza dei redattori della presente relazione che hanno precedentemente definito i criteri di estrazione, così denominati:

a) *anagrafica* (con 5367 records e 16 campi di cui uno è il codice cliente),

b) *peso e misure per giorno* (composto da 13 fogli excel, ciascuno con 12 campi con codice cliente e 65535 recods),

c) *esami* (3782 records per 44 campi con il codice cliente)

I tre files sono stati esaminati ed in alcuni casi modificati ai fini delle analisi statistiche descrittive/ esplorative e le incongruenze sono state comunicate al committente al fine di superare la criticità.

File 1 – Anagrafica -(5367) records ciascuno riferito ad un cliente

Nel file anagrafica sono riportati i seguenti dati

- Codice identificativo
- Sesso (4934 di cui 3979 femmine e 955 maschi)
- Data di nascita (trasformato in età anni compiuti
- Stato di salute (dichiarato)
- Regione di residenza
- Tipo di contratto (il 71% è platinum ed il 29 % Life)
- BMI iniziale (calcolato dal cliente)
- Data inizio forma ideale
- Data attivazione mantenimento
- Peso iniziale in Kg.(comunicato dal cliente)
- Peso a 60 giorni .(comunicato dal cliente)
- Peso a 90 giorni .(comunicato dal cliente)
- Data ultimo peso (può coincidere con quello a 30,60,90 non si può interpretare)
- Ultimo peso
- Polso
- Altezza

File 2- Peso per giorno

Il file è composto 13 fogli excel ciascuno con da 65536 records con le rilevazioni delle seguenti caratteristiche:

- codice cliente,
- data *(in formato gg/mese/anno),*
- fase,
- peso del giorno,
- conta giorni,
- misure antropometriche quali: collo, vita, fianchi, coscia, ginocchio *(non è specificato se Sx o Dx)*, caviglia, petto *(le misure sono fornite dal cliente per cui sono poco attendibili statisticamente in quanto viene introdotta una variabilità soggettiva)*

File 3 – Esami

Il file è composto da 3742 records e riguarda gli esami ematochimici (*andrebbe allineato al file 1)* e sono rilevate le seguenti variabili:

- codice cliente
- azotemia - in quattro occasioni *(mancano le date per cui inutilizzabile)*
- acido urico Mgdl - in quattro occasioni *(mancano le date per cui inutilizzabile)*
- alt - in quattro occasioni *(mancano le date per cui inutilizzabile)*
- ast - in quattro occasioni *(mancano le date per cui inutilizzabile)*
- colesterolo totale Mgdl- in quattro occasioni *(mancano le date per cui inutilizzabile)*
- colesterolo totale Mgdl- in quattro occasioni *(mancano le date per cui inutilizzabile)*
- creatinina Mgdl - in quattro occasioni *(mancano le date per cui inutilizzabile)*
- emoglobina glicata- in quattro occasioni *(mancano le date per cui inutilizzabile)*
- glicemia Mgdl - in quattro occasioni *(mancano le date per cui inutilizzabile)*
- hdl Mgdl - in quattro occasioni *(mancano le date per cui inutilizzabile)*
- ldl Mgdl - in quattro occasioni *(mancano le date per cui inutilizzabile)*
- trigliceridi Mgdl - in quattro occasioni *(mancano le date per cui inutilizzabile)*

1.1-Osservazioni ai tre files

In relazione alle considerazioni effettuate sui tre files ed alle risposte ottenute dal dott. Teta si possono mettere in atto le seguenti azioni:

File 1 anagrafica Peso (5367 records ciascuno con un codice identificativo diverso)

1. alcuni records sono collegabili a clienti che hanno abbandonato la dieta: si consiglia di mantenere il records del cliente che ha abbandonato il programma, per una successiva analisi orientata ad individuare le cause degli abbandoni;
2. inserire dei controlli sui codici cliente per evitare duplicazioni;
3. inserire controlli per evitare errori di immissione di dati relativi alle patologie dichiarate *(si può fornire al cliente una classificazione a priori per evitare errori di immissione)*;
4. nel caso di presenza di patologie individuare una classificazione che permetta di evidenziare quella prevalente o patologie legate ad obiettivi di ricerca successivi legati all'efficacia ed alla sicurezza. Quindi prevedere un campo per ogni patologia ed un campo che indichi il numero di patologie;
5. nel caso di analisi più dettagliate per articoli a carattere scientifico per riviste internazionali è necessario individuare una classificazione condivisibile con altri componenti il gruppo di ricerca;
6. BMI: seguendo le regole per calcolare il peso e l'altezza riportate nel documento di analisi del manuale software Qulik View:
   - calcolare almeno due valori di BMI con due diversi metodi,
   - nel caso di un valore anomalo ripetere il calcolo e confermare il valore,
   - per i casi estremi/anomali mantenerli ma escluderli dal data base di analisi perché inficiano i risultati,
   - per i casi anomali con BMI >40 si potrebbe fare una analisi separata e di confronto tra due popolazioni;
7. Nell' anagrafica per le variabili altezza *(statura)* e circonferenza polso, utilizzare le indicazioni riportate la nel documento di analisi del manuale software Qulik View;
7. Utilizzare i dati mancanti, della data di attivazione mantenimento, per analizzare le cause dei drop-uot;

8.Per le incongruenze tra i valori dei pesi e gli andamenti: vanno eliminati i casi anomali;
9.Mantenere i soggetti che non hanno comunicato il peso per poi analizzarli;
10.Nei casi in cui i clienti non presentano peso a 60, 90gg imputare il peso secondo un criterio (cfr. allegato- osservazioni manuale software QV);
11.Il peso indicato a 60 gg e a 90gg non coincide esattamente con il lag temporale di 60gg e 90gg dall'inizio della FI. Da quanto ci è stato spiegato il termine della FI può essere inferiore a 60gg. Quindi sarebbe utile definire le variabili nel seguente modo: a) peso al termine della FI (ex 60gg), b) peso al termine del la fase di mantenimento (ex 90gg) in questo modo si perde l'informazione dell'influenza del 'tempo' sulla diminuzione del peso. Ci sono 36 casi senza peso finale.

File 2 peso per giorno

1. Evitare di aggiornare le misure del peso nella fase di mantenimento con criteri diversi (ogni 7 gg , ogni 10gg o in base esigenze del cliente). In ogni caso inserire comunque il peso aggiornato in base ad uno dei criteri scelti e utilizzati, al fine di poter coordinare il file 1;
2. colonna C alcuni casi mancanti- ci sono 3571 records con mancata risposta. Trovare un criterio che non imputi i dati giorno per giorno che a nostro avviso non hanno un grande potere informativo, ma scegliere ad es. dati settimanali (ogni sette giorni) o con altro lag temporale che abbia significato o utilizzare un dato medio.
3. esistono 13 Sheet che vanno uniti nella versione più aggiornata di excel.

File 3 esami

1.il numero dei records è 3742 va allineato al file 1;
2.ci sono 16 individui che presentano azotemia nelle 4 occasioni di misurazione (un po' pochi);
3.per estensione ciò può valere anche se non numericamente anche per gli altri esami,
4.dei risultati degli esami può interessare solo il primo ed a 60 gg? cosi abbiamo più casi circa 3600 se invece consideriamo le 3 occasioni 700 casi completi. (i dati si riferiscono all'azotemia ma sono estendibili a tutti gli esami in quanto lo stock di esami è standard);
5.Quali sono gli esami che ti interessano anche in fase prospettica?

Questo file dovrebbe essere ripensato in funzione degli obiettivi che si desidera raggiungere.

Le osservazioni sui tre files sono state discusse con il dott. Teta al fine di utilizzare dei criteri di estrazione dal DB generale, che permettano di costruire un data base attendibile ed utilizzabile direttamente per scopi statistici ed informativi. In linea generale si osserva che non è presente una strategia per costruire un data base finalizzato a degli specifici obiettivi ma si apprezza lo sforzo di acquisire grandi mole di dati per essere successivamente scelti per opportuni scopi.
Le considerazioni sul Manuale Software statistiche **Qlik View**, che è lo strumento per l'archiviazione dei dati reperiti da Bioimis, si trovano *nell'allegato – Osservazioni Manuale Software QV*

1.2 - Considerazioni generali dati Bioimis

I dati presentati da Bioimis da sottoporre per analisi statistiche presentano le seguenti caratteristiche generali:

1. I dati raccolti evidenziano che si tratta di un primo tentativo di costruire una base dati rivolta a individuare le caratteristiche principali dei clienti che si sottopongono alla dieta proposta da 'Accademia alimentare Bioimis';
2. La raccolta ed archiviazione dei dati avviene attraverso un sistema **Qlik View** che evidenzia lo stato del punto 1) e dovrebbe essere aggiornato al fine di ottenere una corretta, coerente ed esaustiva base dati finalizzata a diversi obiettivi (cfr. *allegato – Osservazioni Manuale Software QV)*;
3. La raccolta ed archiviazione dei dati è basata sulla volontà dei clienti a fornire le informazioni richieste quindi si riflette sulla debolezza del DB;
4. Il supporto informatico attuale è una versione datata di excel;
5. I dati devono essere collegati a precise situazioni che non sempre, nella attuale raccolta, vengono rispettate (es. pesi e tempi);
6. Deve essere introdotto un sistema di controllo della qualità che riguardi: 1-la rilevanza (per argomenti e concetti di interesse per le analisi), 2- l'accuratezza (i dati devono essere attendibili), 3- la accessibilità (chiarezza e flessibilità d'uso), 4- la confrontabilità (spazio e tempo, stesse cadenze o criteri); 5- la coerenza (relazioni logiche chiare e rigorose), 6- la completezza (efficacia rispetto al fabbisogno degli utenti);

1.3 - Criteri generali per la costruzione di un DB corretto

In questo contesto si sottolinea l'importanza a) delle modalità operative di selezione dei dati, considerando gli strumenti messi a disposizione dall'informatica, b) di comprendere il valore dell'informazione quantitativa e qualitativa, c) di definire le principali procedure di acquisizione, codifica ed organizzazione dei dati in tabelle finalizzate alla costruzione della base dati da elaborare con metodi statistici. Si ritiene opportuno riportare quanto segue:

1. <u>*I dati statistici*</u> (alcune definizioni): è necessario definire a) l'unità statistica (US) che può essere: individuo, oggetto, collettivo ecc., il cui insieme, costituisce il collettivo oggetto di studio. Nel caso Bioimis, può essere un singolo cliente, un insieme di clienti identificati attraverso una caratteristica *(es. tipologia di patologia)*, b) i **caratteri** o le caratteristiche di interesse che vengono rilevate sulle US che possono assumere le **modalità** numeriche *(in questo caso variabili quantitative es. peso)* oppure attributi *(si parla di variabili qualitative es. tipo di contratto)*, c) le modalità possono assumere le seguenti tipologie: scala numerica ossia un numero, scala ordinale attributo ordinabile *(es. scarso, sufficiente, buono, ottimo*), scala nominale le modalità non sono ordinabili (es sesso : non vi è un ordine naturale o logico tra maschio o femmina). Il dato statistico è il risultato dell'operazione di rilevazione della modalità di risposta. In generale i problemi sono complessi e necessitano di rilevare più dati statistici l'insieme dei quali viene denominato multivariato *(come nel caso Biomis*). I dati statistici devono essere organizzati secondo determinati criteri contenuti nella rilevazione o selezione, nella codifica e ricodifica, controllo dei dati e del trattamento dei dati mancanti.
2. <u>*La codifica e qualità dei dati*</u>: a) per variabili qualitative il codice è legato univocamente alla modalità (es. maschio=1, femmina=2, oppure nord=1, centro=2, sud=3), b) per variabili quantitative il codice numerico corrisponde esattamente al valore numerico che assume la variabile in una US. Ma è possibile comunque, codificare la variabile in classi *(es. età < 30 anni, tra 30 e 50 anni, > di 50 anni.* Il criterio di suddivisione in classi generalmente si basa su classi centrali numerose e classi finali poco numerose come se vi fosse una curva

simmetrica rispetto alla classe centrale, oppure il criterio di eguale numerosità in ciascuna classe. La modalità di scelta del numero di classi e della loro ampiezza influisce sui risultati statistici per cui nei calcoli si utilizzano valori singoli, mentre nelle presentazioni si utilizzano le classi.

3. *dati mancanti e ricodifica*: il trattamento dei dati mancanti avviene in due modi : a) esclusione *(in questo modo si riduce il numero delle osservazioni)* ; b) imputazione *(il dato mancante viene sostituito, utilizzando un apposito criterio che dovrebbe rispettare la regola del dato teorico più vicino)*. I dati possono esser anche ricodificati (per elaborazioni): a) accorpamento e ricodifica su scala binaria (0,1), suddivisione in classi.
4. *La qualità dei dati* in genere si seguono i seguenti aspetti: Rilevanza (argomenti e concetti di interesse), Accuratezza (stime attendibili), Tempestività (gap ridotto tra produzione del dato e fruizione dello stesso), Accessibilità (chiarezza e flessibilità dell'uso), Confrontabilità (spazio-tempo), Coerenza (relazioni logiche chiare e rigorose), Completezza (efficacia rispetto al fabbisogno dell'utente).
5. *Le fonti di errore*: errori non campionari (imprecisioni codifica, rilevazione, imputazione, errori campionari (indagine campionaria).
6. *L'organizzazione dei dati in tabelle*: di intensità, di valori medi, booleane, di punteggi, di preferenze, di ranghi, ed anche le matrici di uso comune per la individuazione di relazioni multivariate: similarità/dissimilarità, intensità di flussi in matrici di scambio, esistenza di una relazione (booleana). La matrice dei dati in Excel tabella è un insieme di celle disposte secondo righe (identificate da numeri) e colonne identificate da lettere ed è un foglio di lavoro, la cartella di lavoro è costituita da un insieme di fogli di lavoro. In senso statistico in una generica tabella si individuano le colonne che sono intese come variabili statistiche (valori-risposte) e di righe che rappresentano l'insieme delle risposte, date da ciascun individuo ad ogni variabile.
7. *La 'trasformazione' dei dati*. Percentuali per riga, per colonna, centrati, standardizzati, ridotti al fine di ottenere: indicatori statistici, rapporti statistici dati pro-capite o di densità o numeri indice.
8. *Strategie di ricerca e preparazione del dato*, i fatti (le osservazioni empiricamente verificabili, la teoria, i concetti le ipotesi di lavoro.
9. *I dati ed il valore dell'informazione in azienda*: l'aumento del volume dei dati operazionali in azienda rende necessario il supporto informatico per processi decisionali (nel caso Bioimis anche la scelta della tipologia di dieta nel caso fosse necessario). Il 'magazzino dei dati viene anche denominato *data warehous*. Il *data warehouse*, quindi, descrive il processo di acquisizione, trasformazione e distribuzione di informazioni presenti all'interno o all'esterno delle aziende come supporto ai *decision maker*. Esso si differenzia in modo sostanziale dai normali sistemi gestionali che, al contrario, hanno il compito di automatizzare le operazioni di *routine*. Informazione che è l'insieme di dati in grado di cambiare/arricchire le nostre conoscenze, la conoscenza che consiste nella trasformazione di informazione in valore. Informazione è necessaria per pianificare e controllare attività con efficacia. Essa viene trasformata e trasmessa ai sistemi informativi per essere parte integrante del data warehous. L'informazione è una risorsa aziendale e come tale ha un valore ed un costo relativo. Determinante che l'informazione sia corretta e contenga elementi per la comprensione ed i limiti delle informazioni. Come diceva M.E.Porter (Harvard Business School) - dare l'informazione giusta alla persona giusta, nel momento giusto per prendere la giusta decisione-.

10. *Esigenze in azienda*: le frasi comuni: abbiamo montagne di dati, vogliamo tagliare i dati a fette in ogni modo, mostratemi ciò che è importante, tutti sanno che alcuni dati sono corretti. I data base sono una grossa risorsa potenziale che solo se utilizzata e costruita correttamente può dare benefici sostanziali.
11. *Automatizzare l'estrazione* della conoscenza attraverso tecniche di data minig, che permettono di estrarre pattern di dati tramite algoritmi che individuano le associazioni nascoste tra le informazioni e le rendono visibili. A seguire i sistemi di supporto alle decisioni che permettono di definire delle regole per le azioni da mettere in atto.
12. *Il data warehouse*: organizzato in quattro livelli di architettura: i) trasformazione dei dati (acquisire dati e validarli); ii) preparazione e stoccaggio dati (fornisce i dati agli utenti e alle applicazioni analitiche); iii) interpretazione e analisi dei dati (livello elevato di valore aggiunto che presiede alla trasformazione dei dati in informazioni aventi valore strategico); iv) presentazione dei dati (basso valore aggiunto in genere estetico per gli utenti):

Al fine di facilitare la gestione di obiettivi operativi è opportuno progettare una piattaforma di sistema WD con caratteristiche: facilità di accesso alle informazione, gestione delle versioni storiche dei dati, visione multidimensionale dei dati, capacità di costruire scenari futuri.

1.4 – Considerazioni finali sul DB esaminato e progressione analisi statistiche:

- Per avere una data base che permetta, la estrazione di campioni statisticamente elaborabili e giustificabili, è necessario *specificare le regole di reclutamento dei clienti* che sono la discriminate per le informazioni che si possono estrarre da un qualsiasi campione del DB e che comunque vanno specificate nel caso di lavori scientifici sia per riviste scientifiche nazionali ed internazionali;
- Utilizzare un sistema di archiviazione dati che contenga una serie di controlli ed avvisi (warnig) per evitare: mancate risposte e dati anomali che richiedono una conferma da parte dell'operatore (alcune indicazioni sono contenute nell'allegato "Osservazioni manuale software QV " utilizzato da Bioimis;
- Arricchire il DB con nuove variabili quali ad es. (motivazionali sulla scelta del programma alimentare di Bioimis;
- Assenza per molti casi della associazione dato del peso e data di rilevazione;
- Nel caso di analisi riguardanti la sicurezza del programma alimentare saranno necessarie specifiche analisi anche considerando un opportuno campione di controllo;
- Per completare il quadro generale del DB di Bioimis sarebbe opportuno integrare gli attuali dati con le motivazioni sulla scelta del programma alimentare oltre che ad obiettivi di raggiungimento. Sarebbe opportuno integrare il data base utilizzando un programma che generi da QV, il codice cliente e data di inizio del programma alimentare con le seguenti variabili correlate al programma alimentare: a) motivo della scelta di seguire una dieta: *estetico, di salute, psicologico, altro (specificare), b)* perché hai scelto Bioimis?: *ho provato altre diete e non hanno funzionato, mi sembra un approccio non troppo invasivo/ impegnativo, la tipologia di scelta di Bioimis di seguire il cliente con assiduità, altro (specificare); c)* come sei venuto a conoscenza di Bioimis: *passa parola, attività dell'agente di zona, da programmi televisivi, da giornali da altre fonti (indicare); d)* il motivo dell'uscita dalla forma ideale: *raggiungimento obiettivo, uscita volontaria, uscita consigliata da Bioimis*; e) quante volte è entrato/uscito dalla FI prima di aver raggiunto l'obiettivo ed entrato in mantenimento: 1 volta, 2 volte, 3 volte , più di 3 volte ; f) quante volte è entrato/uscito dalla fase di mantenimento prima di aver raggiunto l'obiettivo in forma per sempre: 1 volta, 2 volte, 3 volte , più di 3 volte ; g) definire una variabile

codificata che sia inerente all'obiettivo del cliente (es . perdere X chilogrammi); h) valutazione del cliente in dieta: al momento attuale (attenzione alla fase) ritieni di aver ottenuto un risultato rispetto agli obiettivi prefissati: *non soddisfacente, soddisfacente , buono , ottimo;*

- Sapere quale tipologia di dieta sia più efficace correlata alla tipologia di cliente. in modo da reclutare clienti che, per caratteristiche comuni hanno avuto maggiore efficacia nella dieta. Ma anche rispetto ad una classificazione iniziale rispetto al BMI.
- Per quanto attiene al documento 'Analisi statistiche programma alimentare Bioimis', che si sintetizza in due obiettivi fondamentali: Efficacia e Sicurezza, va evidenziato che il punto 1 è stato raggiunto dopo una attenta analisi dei dati che è stata la parte più laboriosa e necessaria per effettuare le analisi statistiche. In particolare l'analisi dei dati ha riguardato le mancate risposte, le variabili relative alla massa corporea ed agli esami antropometrici. Per quanto attiene alla sicurezza va impostata una nuova fase di lavoro per obiettivi in modo da predisporre i dati in modo corretto per le analisi statistiche anche attraverso un campione di controllo.

2 -Le strategie di analisi

Dopo aver esaminato il file dei dati e constato che il campione consegnato da Bioimis presentava: delle incongruenze di varia natura, mancate risposte al peso ed alle variabili ematochimiche ed a seguito della interlocuzione con il dott. L. Teta, che ha provveduto colmare le carenze segnalate, si è stabilita la seguente strategia di analisi:

a) dare al committente una descrizione generale sulle variabili del file anagrafica;

b) classificare le mancate risposte collegate alle variabili più significative attraverso la *metodologia della segmentazione binaria* (alberi di classificazione CARTS) utilizzando l'intero campione di 5367 individui;

c) Vista la notevole mole di m.r. oltre ad individuare i clienti con propensione a fornire mancate risposte si è proceduto ad associare alle variabili più significative la probabilità di fornire mancate risposte utilizzando l'intero campione con 5367 individui. Ciò al fine di fornire indicazioni per predisporre azioni rivolte al miglioramento della qualità dei dati da parte di Bioimis;

d) prima analisi sull'efficacia effettuata sul campione senza mancate risposte composto da 2684 individui. Si è applicata la metodologia della *cluster analisys* al fine di classificare i clienti in gruppi omogenei tenendo conto di tutte la variabili anagrafiche;

e) L'ultimo studio, a completamento della fase osservazionale, è attuata analizzando anche il fattore tempo, attraverso una analisi di tipo longitudinale rivolta all'efficacia della dieta (Biomis) a ridurre il peso dei soggetti che hanno intrapreso il programma alimentare proposto .

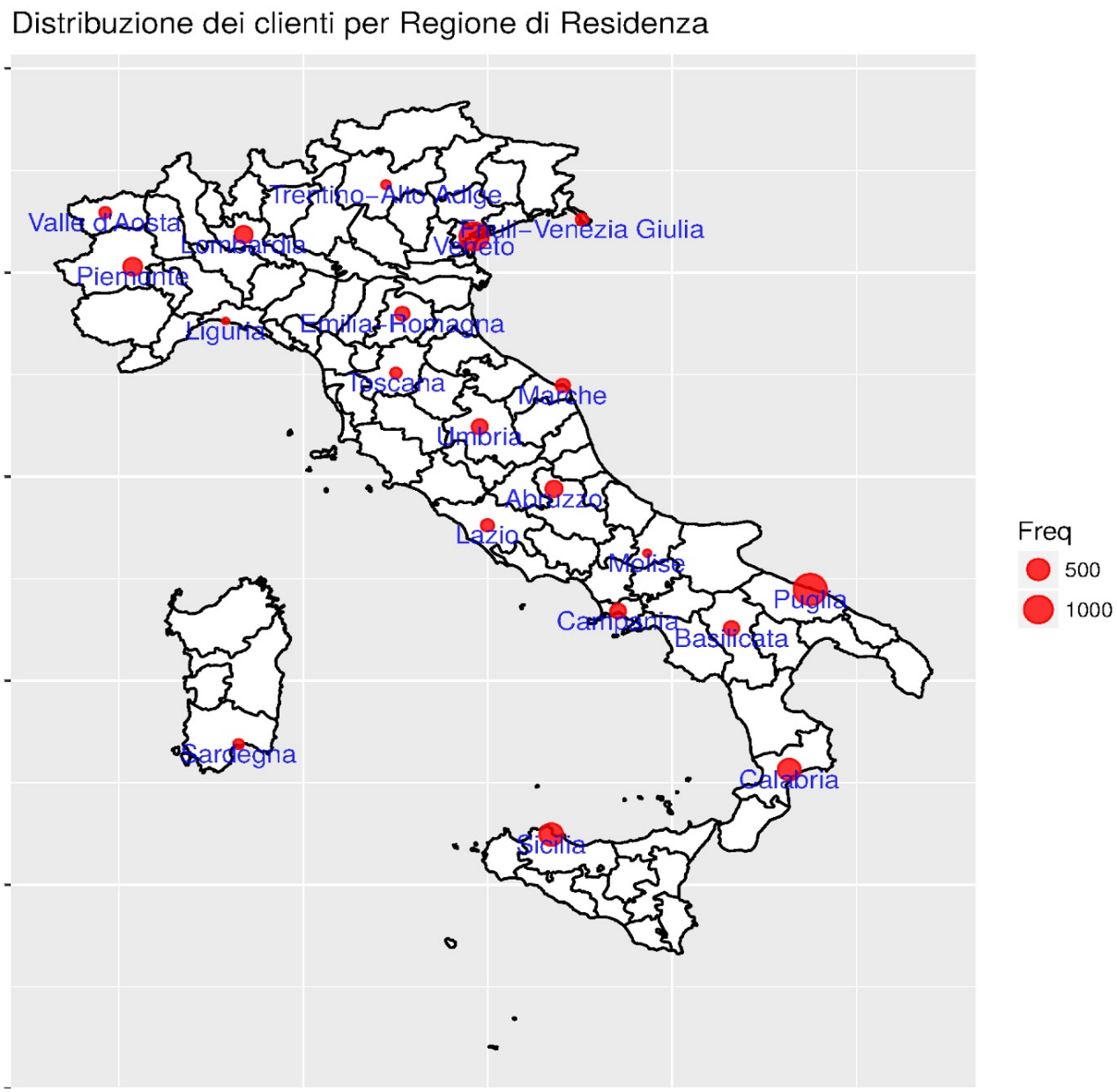
3 – Descrizione dei dati

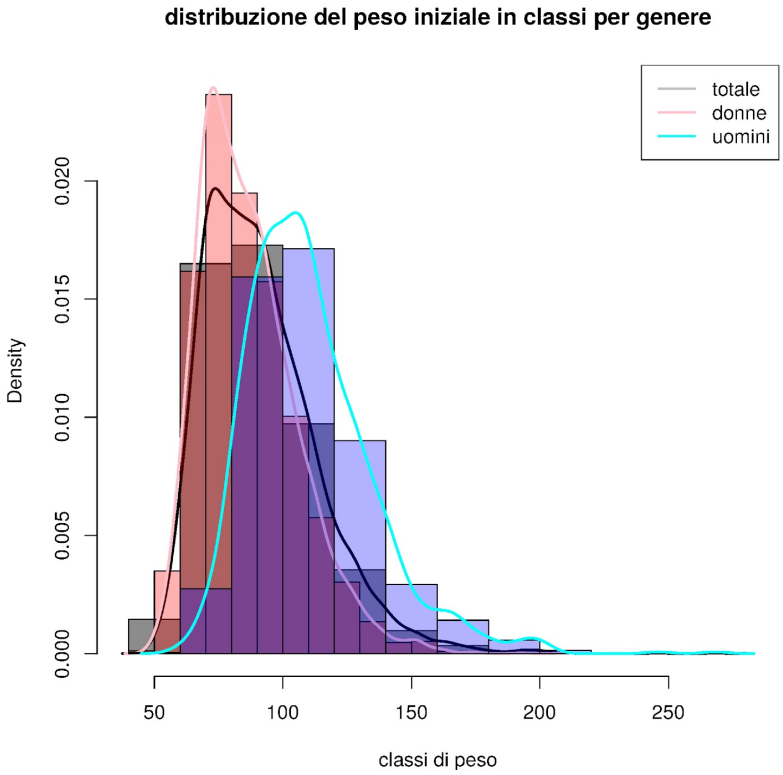
I dati del campione in esame senza i casi che presentano mancate risposte si riducono da 5367 individui a 2684. Le caratteristiche principali sono le seguenti:

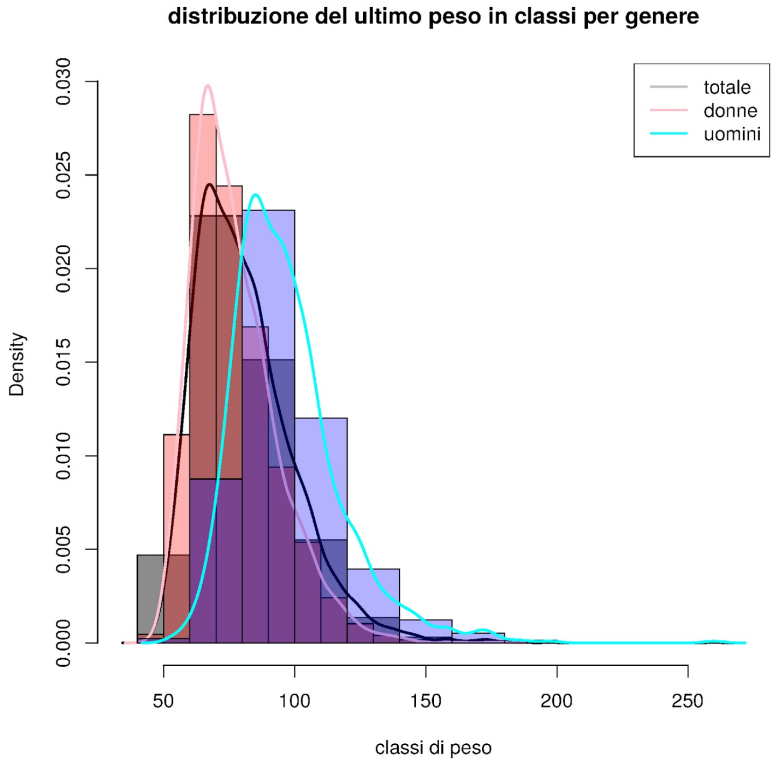
a) il 79% è costituito da 'clienti di sesso femminile';
b) il 66% non ha dichiarato patologie;
c) il 78% ha un contratto platinum;
d) il 45% è residente al sud ed il 32% al nord.

Per quanto attiene alle variabili in analisi: BMI iniziale medio di 34,1; peso iniziale medio 93,85 kg.; peso a 60 gg medio 83,52 kg.; peso a 90 gg medio 81,55 kg.; ultimo peso medio 81,92 kg.; durata media in forma ideale 74,01 giorni.

La distribuzione per regione è evidenziata nel cartogramma sottostante. Le distribuzioni del peso riportate nel seguito, pur nella diversità delle distribuzioni per uomini e donne rispetto alla media, evidenziano una diminuzione del peso medio.




distribuzione del peso iniziale in classi per genere
totale
donne
uomini
Density
0.000
0.005
0.010
0.015
0.020
50
100
150
200
250
classi di peso



4 – Segmentazione binaria

Principali aspetti metodologici

La prima informazione che si è ritenuto acquisire, vista la tipologia dei dati forniti da Bioimis, consiste nell'individuare le caratteristiche principali degli individui, sottoposti alla dieta alimentare, che non hanno comunicato il proprio peso l'ingresso nella fase forma ideale (IFI) o ai momenti tipici richiesti: dopo 30gg, 60gg, 90gg .

Si è optato di utilizzare la metodologia degli 'alberi di classificazione ' (CARTS: Classification & Regression Trees). Questa metodologia ha la particolarità di essere classificata, tra i metodi non parametrici di classificazione e regressione che utilizzano una struttura ad albero di facile interpretazione e applicazione a scopi decisionali. *(Breiman L., Friedman J.H.,Olshen R., Stone C.J. "Classification and Regression Trees" 1984).*

Obiettivo può essere esplorativo o predittivo che comunque si traduce nel partizionare 'ricorsivamente' un collettivo di unità statistiche (tutti gli individui del campione, nel nostro caso) in sottogruppi binari, in modo tale che le unità statistiche siano il più possibilmente omogenee all'interno di ogni sottogruppo. L'omogeneità è misurata come mutabilità o variabilità di una variabile target (risposta), che può esse quantitativa o qualitativa.

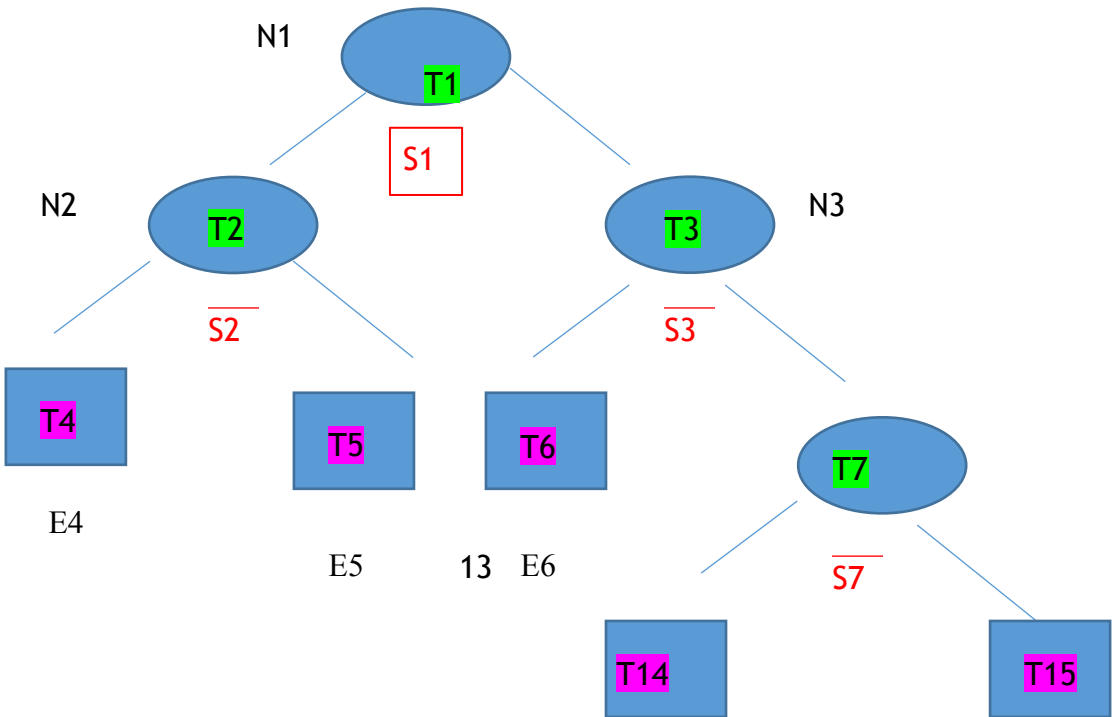
Il processo di classificazione è effettuato sulla base della migliore divisione dicotomica, tra tutte quelle indotte dalle possibili divisioni delle modalità di ciascun predittore dove, per predittore si intende una variabile che entra nell'analisi ai fini esplicativi della variabile target (in questo caso risposta no/si).

Il risultato è una struttura ad albero che consente di comprendere la gerarchia di importanza dei predittori nella spiegazione della variabile target, stabilendo per ogni percorso dell'albero, dal nodo iniziale (padre) ai suoi nodi terminali, le interazioni tra predittori nella definizione dei gruppi finali e della loro composizione.
Il risultato è una serie di regole di decisione o previsione per attribuire: ad una classe (variabile target qualitativa) o un valore numerico (variabile target quantitativa), una unità statistica sulla base delle sole misurazioni di un insieme di predittori. L'obiettivo predittivo nella presente analisi sarà sviluppato in seguito con un modello logistico che sarà illustrato in un successivo paragrafo. Anche nel caso di scopo predittivo, la tipologia di dati è costituita da un insieme di variabili esplicative (predittori) ed una variabile (target di risposta) ambedue le tipologie di dati possono essere sia qualitative che quantitative. In genere si distingue un campione di apprendimento che costituisce l'albero esplorativo dal campione test utile per identificare l'albero delle decisioni. Nel grafico seguente viene esposta una tipologia di albero di classificazione.
Illustrazione dell'albero:

- Nodi intermedi sono : T1, T2, T3, T7;
- Nodi terminali sono: T4, T5, T6, T14, T15;
- Gli Split (punto in cui un nodo si spacca in due sotto nodi figli) sono: S1, S2, S3, S7 sulla base di una variabile discriminante;

Le 'Etichette' (che descrivono i nodi terminali sono delineate dalle variabili del set di dati) sono: E4, E5, E6 E14 E15;

-



E14 E15

La numerazione dei nodi tiene conto di una semplice regola pratica: il nodo (T) genera il nodo figlio di sinistra (2T) ed il nodo di destra (2T+1). In questo modo, a partire dal nodo terminale è possibile risalire al nodo radice in maniera univoca.
L'insieme di split candidati : definisce, per ogni nodo intermedio l'insieme delle domande binarie, ossia l'insieme delle divisioni ammissibili, detti split, del gruppo di unità presenti nel nodo. Ogni split di unità statistiche è generato dalla bipartizione delle modalità di un predittore in due gruppi. L'insieme degli split è il totale di possibili split generati da ciascun predittore.
La scelta del migliore split è un problema computazionalmente difficile dato l'alto numero di combinazioni in cui si possono bipartire le modalità di variabili qualitative politomiche. Tra le varie soluzione implementative si è scelto di far riferimento al metodo CART (Classification and Regression Trees) .
Obiettivo specifico è quello di generare nodi figli più puri, cioè più omogenei al loro interno, dei nodi padri rispetto al carattere oggetto di classificazione (variabile target). In genere nell'ambito della classificazione i nodi figli sono più omogenei rispetto ai nodi padri e gli individui che vi appartengono sono in gruppi più piccoli.
La misura dell'impurità del nodo è di solito definita dall'eterogeneità nel caso di variabile target qualitativa. Una Eterogeneità minima significa che tutte le unità statistiche sono in una sola classe di risposta e quindi il gruppo di unità statistiche è perfettamente omogeneo rispetto alla variabile target. Eterogeneità massima si ottiene invece nel caso in cui le unità statistiche si equi-distribuiscono perfettamente tra le classi di risposta così che non esiste una sola moda della variabile target.
Tra i vari possibili criteri di arresto alla bipartizione due sono particolarmente degni di nota:

- Numerosità minima in un nodo: ossia si dichiara terminale un nodo che ha meno di una % prefissata di unità statistiche del collettivo o un nodo in cui tutte le unità sono omogenee rispetto alla variabile risposta;
- test statistico che valuta, sulla base di un campione indipendente la diminuzione di impurità ottenuta effettuando lo split

Il '*Prunig*' si basa sulla costruzione di un albero totalmente espanso e progressivamente si tagliano i rami deboli sulla base di un coefficiente costo/beneficio opportunamente definito. In questo modo si perviene ad una sequenza innestata di sottoalberi dai quali scegliere l'albero delle decisioni per nuove unità. Sia il *pruning* che la valutazione della performance degli alberi di classificazione e regressione a fini predittivi viene fatta attraverso una tecnica nota come cross-validation, in cui le unità del collettivo non vengono tutte utilizzate per costruire l'albero, ma alcune vengono casualmente selezionate come campione indipendente su cui testare l'albero ottenuto.

L'applicazione della tecnica della segmentazione al set di dati

Si premette che l'analisi è stata effettuata su tutto il set di dati 'anagrafica' che comprende **5637** casi. Questa scelta è dettata dall'ipotesi di lavoro che considera i dati del peso, che costituiscono

l'obiettivo della dieta, come una variabile risposta. Nel set iniziale si trovano molti record con la 'stringa' dei pesi :"peso iniziale", "peso a 30 giorni", "peso a 60 giorni", "peso a 90 giorni" e "ultimo peso" non sempre completa. Ciò evidenzia un 'baco' nel reperimento dei dati od una mancata volontà di comunicare il peso, da parte del cliente, alle date richieste per differenti motivi. Questo fattore produce set di dati non sempre utilizzabili per le analisi statistiche che prevedono sempre delle misurazioni al fine di dare valore ai risultati. E' ben noto che la qualità di un 'minestrone' dipende dagli ingredienti e così vale per la consistenza e validità dei risultati statistici. Una prima informazione che si può estrarre dai dati grezzi è proprio legata alla propensione dei propri clienti a fornire il valore del peso richiesto. Pertanto l'approccio seguito, ricerca la possibilità di abbandonare, per sempre o per un certo periodo la dieta proposta. A tal riguardo si è applicata la tecnica della segmentazione su set di dati distinti: a) il primo riferito al numero di occasioni che il cliente fornisce i dati del peso considerando l'intero set di clienti e tutte le variabili presenti nel dataset "anagrafica", b) il secondo deriva dal primo set di dati avendo eliminato la variabile "durata di FI". Questa scelta deriva dalla considerazione che questa variabile (durata FI) abbia un effetto sulla comunicazione del proprio peso alle date consigliate. Di seguito è riportato l'albero finale di classificazione considerando tutte le variabili del file anagrafica. Il criterio di lettura unico per tutti i nodi ed è specificato nel seguito per il nodo padre indicato con T1.

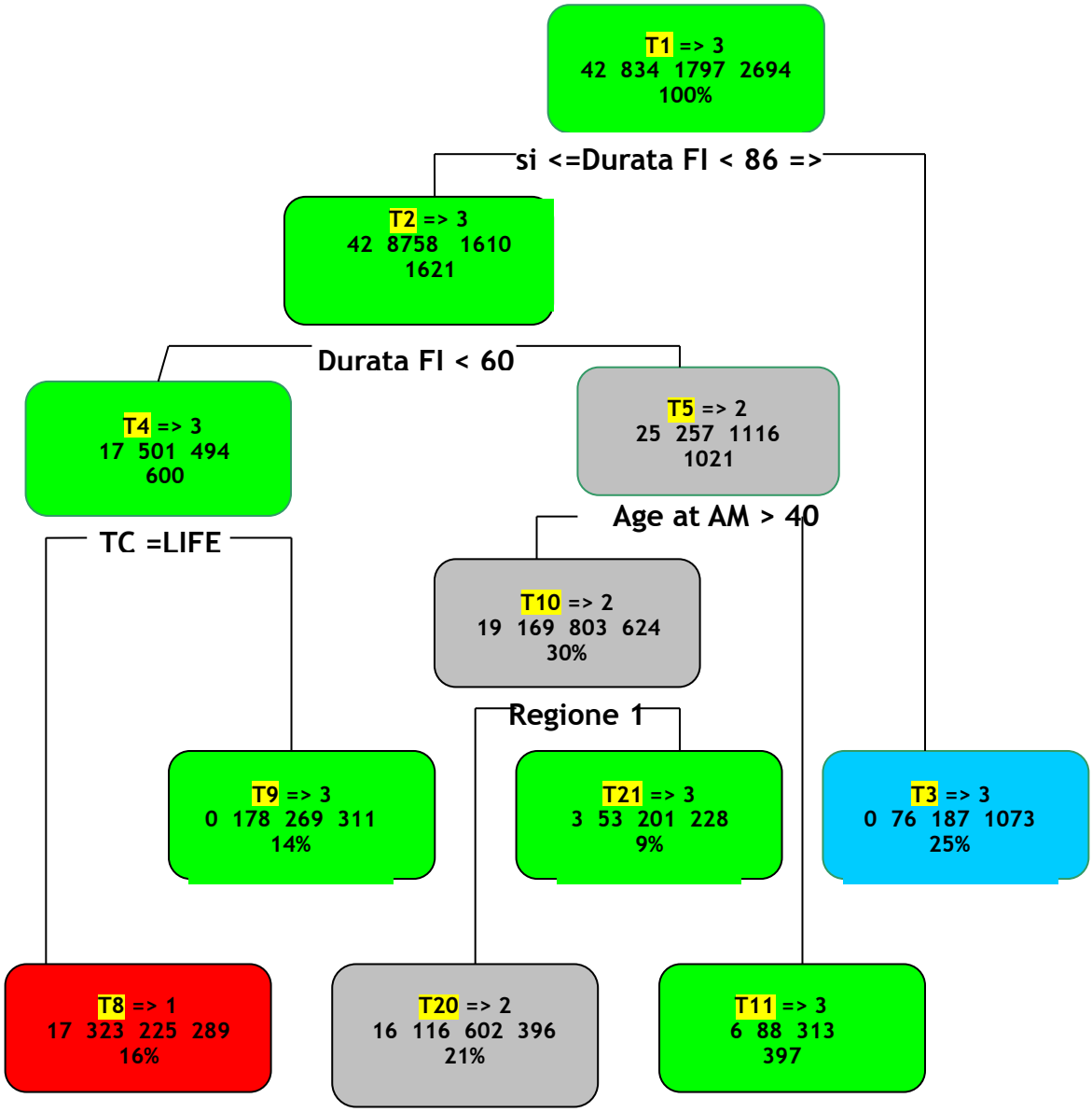
Il nodo padre dell'albero costituito da 5637 individui, è suddiviso in funzione del numero di volte che un cliente comunica il proprio peso indipendentemente dall'occasione richiesta. Ipotesi di lavoro:"la propensione a fornire i dati sul proprio peso". L'ipotesi deriva dalla seguente considerazione: il cliente accetta di sottoporsi ad una dieta per diminuire il proprio peso.

I 5647 clienti appartenenti sono così classificati in funzione del numero delle mancate comunicazione del peso: 2694 unità ha comunicato il proprio peso in tre occasioni (50% del nodo padre) e costituisce la modalità più frequente; 1797 unità hanno comunicato il proprio peso in due occasioni (33% del nodo padre); 834 unità hanno comunicato il proprio peso solo in una occasione;

42 unità non hanno mai comunicato il proprio peso (0,7% del nodo padre) .Questi ultimi sono soggetti il cui percorso di dieta sarà destinato molto probabilmente a non concludersi.

Su questo contingente si è costruito un albero sempre legato agli abbandoni (temporanei o definitivi che hanno generato mancate risposte del peso).

Albero del numero di risposte al peso per cliente



L'albero ottenuto evidenzia una prima partizione (split S1) in due gruppi dove **la variabile discriminante è la durata di permanenza in FI:**

1. **1.mo Gruppo** (nodo intermedio e nodo terminale T3): 1336 individui presentano una durata in **FI > di 86** giorni che sono il 25% dell'intero campione di dati (nodo padre). Questo gruppo essendo un *nodo terminale* (T3), nel processo di classificazione, è utile descriverlo:
   a) in prevalenza (1073 clienti) hanno comunicato 3 osservazioni del proprio peso (80% del nodo terminale),
   b) i rimanenti 187 (14% del nodo terminale) hanno comunicato due volte il proprio peso**,**
   c) la parte rimanente ha comunicato solo un peso in 76 casi (6% del nodo terminale),
   d) tutti i clienti in questo nodo hanno comunicato almeno un peso

   Descrizione: Questo gruppo (nodo T3) presenta soggetti con maggiore coerenza rispetto al programma alimentare e sono circa il 25% del nodo padre (T1) tra essi i più scrupolosi (1073 unità) costituiscono il 19% dell'intero campione
2. **2.do Gruppo** (nodo intermedio T2): 4031 individui presentano una durata in **FI < di 86** giorni. E costituiscono circa il 75% dell'intero campione (nodo padre T1). Questo nodo intermedio è così costituito:
   a) la maggior parte ha comunicato 3 osservazioni del peso,1621 soggetti circa il 40% del nodo intermedio e il 29% del nodo padre),
   b) appena sotto, come incidenza, le unità che hanno comunicato due osservazioni del peso 1610 (circa il 39% del nodo intremedioT2 e 28% del nodo padre T1),
   c) 758 soggetti hanno comunicato un peso (circa 18% del nodo intermedio T2 ed il 13% del nodo padre T1),
   d) permangono le 42 unità iniziali che non hanno mai comunicato il loro peso (1% di T2 e 0,7% del nodo padre).

   Questo secondo gruppo si scinde ulteriormente in due nodi intermedi T4 e T5 sempre in funzione della **variabile discriminante - durata di permanenza in FI:**
   - T4 con 1612 soggetti con **con IF < di 60** gg (60 % del secondo gruppo T2 ed il 29% del gruppo inziale T1).
     - ✓ Questo nodo intermedio si suddivide in due nodi terminali T8 e T9 in base alla variabile discriminante tipo di contratto: Life (T8) e Non Life (T9) che costituiscono due nodi terminali.

     Descrizione: il nodo T8 costituisce il peggior gruppo di clienti in riferimento alla comunicazione del loro peso. E' caratterizzato dalla prevalenza di individui con una sola risposta. Il nodo T9 invece non presenta mancate risposte di peso e la maggior parte comunica tre volte il proprio peso anche se il 35% lo comunica 2 volte.
   - T5 con 2419 soggetti **con IF > di 60 gg e < di 86 gg** che sono il 45% del gruppo iniziale. Questo nodo intermedio si suddivide in altri due sotto nodi T10 e T11 che diviene nodo finale.

     Il nodo finale T11 con 804 unità (33% di T5 ed il 14% di T1) ha una prevalenza di tre osservazioni ma anche un consistente numero di unità con due osservazioni
   - T10 con 1615 unità (costituisce il 30% del nodo T5) si suddivide, in base alla regione di residenza in due nodi finali T20, con 1130 unità residenti in: Abruzzo, Basilicata, Calabria, Emilia Romagna, Lombardia, Piemonte, Puglia, Sardegna,

Sicilia, Toscana, Trentino alto Adige , e T21 con 488 unità residenti in Veneto, Friuli, Liguria, Marche, Umbria, Lazio, Campania,

- Descrizione: **il noto finale T20 presenta una alta frequenza di due risposte al peso ed anche di mancate risposte quindi non costituisce un gruppo di soggetti che seguono le indicazioni sulla comunicazione del peso. Il nodo finale T21 composto da 485 unità presenta una prevalenza di tre osservazioni di peso (80% del nodo intermedio T8 ma anche una elevata percentuale di due osservazioni del peso (41%).**

In conclusione si individuano 6 nodi terminali (gruppi) che in funzione del numero di mancate risposte alle varie occasioni possono essere ordinati dal peggiore gruppo al migliore gruppo:
1.mo nodo terminale T8 (peggiore: prevalenza di una osservazione e massimo numero di mancate risposte) è caratterizzato dalle variabili discriminanti: durata FI minore di 60 gg e tipo di contratto Life;
2.do nodo terminale T20, (prevalenza di due osservazioni e alto numero di mancate risposte 16) caratterizzato da una durata di FI tra 60 e 86 gg , un'età maggiore di 40 anni[1] e residente in Abruzzo, Basilicata, Calabria, Emilia Romagna, Lombardia, Piemonte, Puglia, Sardegna, Sicilia, Toscana, Trentino alto Adige;
3.zo nodo terminale T9 è caratterizzato dalle variabili discriminanti: durata FI minore di 60gg e tipo di contratto Platinum;
4.to nodo terminale T21 è caratterizzato dalle variabili discriminanti: durata FI tra 60 gg e 86 gg età maggiore di 40 anni e residente nelle regioni Veneto, Friuli, Liguria, Marche, Umbria, Lazio, Campania;
5.to nodo terminale T11, è caratterizzato dalle variabili discriminanti: durata FI tra 60 gg e 86 gg età minore di 40 anni;
6.to nodo terminale T3 (il migliore caratterizzato dalla variabile discriminate durata in FI maggiore di 86 gg. ed assenza di mancate risposte)

Sintesi della segmentazione ad albero in funzione delle variabili caratteristiche

gruppo	numerosità	Prevalenza risposte al peso e loro incidenza nel gruppo	Peso del gruppo	!a Variabile caratteristica	2a Variabile caratteristica	3a Variabile caratteristica
T3	1336	3 (80%)	25%	FI > 86gg	-	-
T8	854	1 (38%)	16%	FI < 60gg	Life	
T9	758	3 (41%)	14%	FI < 60gg	Platinum	
T20	1130	2 (53%)	21%	FI 60-86 gg	Età > 40 anni	Regione 1(*)
T21	483	3 (47%)	9%	FI 60-86 gg	Età > 40 anni	Regione 2(**)
T11	804	3 (49%)	15%	FI 60-86 gg	Età < 40 anni	-

Nota: Per età è stata considerata l'età all'attivazione del mantenimento (AM)
(*)Regione 1 : Abruzzo, Basilicata, Calabria, Emilia Romagna, Lombardia, Piemonte, Puglia, Sardegna, Sicilia, Toscana, Trentino alto Adige

(**)Regione 2: Veneto, Friuli, Liguria, Marche, Umbria, Lazio, Campania,

Descrizione dei gruppi

1 gruppo nodo terminale (T3) di 1336 soggetti (25%) del campione caratterizzato da una prevalenza di tre risposte al peso, con durata in FI > di 86gg

2 gruppo nodo terminale (T8) di 854 unità pari al 16% del campione, caratterizzato da una prevalenza di una risposta al peso, una permanenza in FI < di 60gg e tipo di contratto Life;

3 gruppo nodo terminale (T9) di 758 unità pari al 14% del campione, caratterizzato da una prevalenza di tre risposte al peso, permanenza in FI < di 60gg, contratto Platinum ;

4 gruppo nodo terminale (T20) di 1130 unità pari al 21% del campione, caratterizzato da una prevalenza di due risposte al peso, permanenza in FI tra 60 e 86gg, età ingresso nella fase di mantenimento >40 anni e residenti nelle regioni di tipo 1 (cfr*);

5 gruppo nodo terminale (T21) di 483 unità pari al 9% del campione, caratterizzato da una prevalenza di tre risposte al peso, permanenza in FI tra 60 e 86gg, età ingresso nella fase di mantenimento > 40 anni e residenti nelle regioni di tipo 2 (cfr**);

6 gruppo nodo terminale (T11) di 1804 unità pari al 15% del campione, caratterizzato da una prevalenza di tre risposte al peso, permanenza in Fi tra 60 e 86gg, età ingresso nella fase di mantenimento <40 anni;

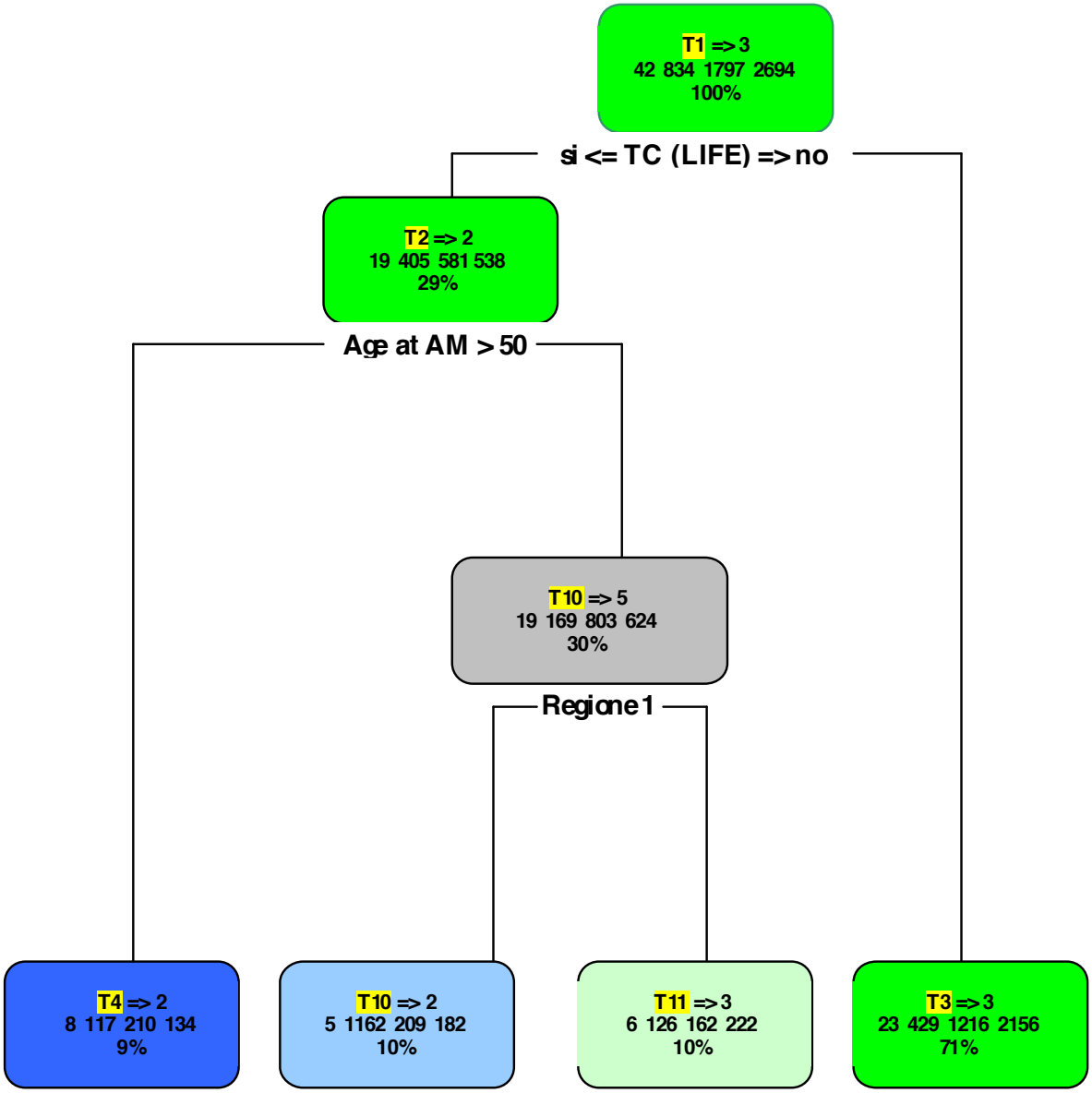
In conclusione le variabili che intervengono nel processo di bipartizione dei clienti sono, in ordine di importanza:

1. **Durata FI:** (>86gg i più disposti a fornire il peso e minore di 86gg che presentano altre variabili caratteristiche)
2. **Durata FI: (tra 60 e 86 gg si-no)**
3. **Age at AM** (age < di 40 anni, i più giovani, e age > di 40 anni più anziani);
4. **Tipo di contratto** (Platinum e Life)
5. **Regione di residenza**

Considerando che la prima variabile discriminante nell'analisi precedente era la durata in FI si è affinata la costruzione dell'albero eliminando dall'analisi la durata della FI. Tale scelta è dettata dal fatto che la variabile "durata della FI" è fortemente connessa con il numero di misurazioni riportate e non necessariamente è lei ad influenzare il numero di osservazioni ma può essere il numero di osservazioni ad influenzare la durata della FI, dato che un soggetto che è presente per tutto il percorso della forma ideale probabilmente presenterà più misurazioni di un soggetto che decide di uscire dallo studio dopo una sola misurazione. A tal scopo si è scelto quindi di rimuovere tale variabile da quelle selezionabili nello *splitting* dei nodi.

Il risultato ha prodotto la seguente albero che evidenzia la presenza di 4 gruppi finali facilmente interpretabili.

Albero del numero di occasioni per cliente senza durata FI



Descrizione dell'albero senza considerare la durata della FI

Sempre seguendo il criterio di lettura già descritto, l'albero ottenuto evidenzia una prima partizione (split1) del campione in esame (nodo padre) in due gruppi in base alla v**ariabile discriminante tipo di contratto**: che è così composto: 42 (persone con nessun dato sul peso) , 834 (persone con 1 dato sul peso), 797 (persone con 2 dati sul peso), 2694 (persone con 3 datai sul peso). L'albero è stato costruito tenendo conto delle variabili: peso (variabile risposta), tipo di contratto, età di ingresso in mantenimento, regione di residenza.

Le persone che hanno comunicato il peso in almeno 3 occasioni tenendo conto del tipo di contratto, età, regione sono stati classificati non considerando la durata della forma ideale.

I soggetti rispetto al numero di osservazioni del peso con il massimo di osservazioni 3 sono così classificati: 42 (nessun dato sul peso) , 834 (1 dato sul peso), 797 (2 dati sul peso), 2694 (3 datai sul peso). Questo costituisce il nodo padre che di *splitta* nel seguente modo:

- **1.mo gruppo (nodo terminale T3**) – (71% del campione) 3824 clienti caratterizzati da contratto **Platinum.** Essi hanno le seguenti caratteristiche: non hanno mai comunicato il peso 23 individui (1%), hanno comunicato il peso una volta 429 individui (11%), hanno comunicato il peso due volte 1216 individui (32%), hanno comunicato il peso tre volte 2156 individui (56%);

  Descrizione: Questo gruppo (nodo finale T3) presenta soggetti con maggiore coerenza rispetto al programma alimentare e sono circa il 71% del nodo padre (T1) tra essi i più scrupolosi sono 2156 unità che costituiscono il 38% dell'intero campione.

- **2.do gruppo** *(nodo intermedio T2)* - (29% del campione) 1673 clienti caratterizzati da contratto **Life**. Essi hanno le seguenti caratteristiche: non hanno mai comunicato il peso 19 individui (1%), hanno comunicato il peso per una volta 405 individui (11%), hanno comunicato il peso due volte 581 individui (32%), hanno comunicato il peso 3 volte 538 individui (56%).

  Questo gruppo intermedio si suddivide ulteriormente in base **all'età di ingresso nella fase di mantenimento (Age at AM)** nei seguenti due gruppi:

  - **gruppo nodo terminale (T4)** caratterizzato da clienti - **con Age ad AM > di 50 anni** - che sono 469 (il 30% del nodo precedente) caratterizzati da: nessun peso fornito 8 (2%), individui che hanno fornito il peso una sola volta 117 (25%), individui che hanno fornito il peso due volte 210 (45%) e individui che hanno fornito il peso per 3 volte (28%);

    Descrizione: il nodo T4 è di piccola entità con clienti più anziani ma non è compatto in quanto sono presenti tutte le modalità
  - *gruppo nodo intermedio T5* – caratterizzato da clienti - **con Age ad AM < di 50 anni** – che sono caratterizzati da: non hanno comunicato il peso 11 (1%), hanno comunicato il peso una volta 288 (15%), hanno comunicato il peso due volte 371 (63%), hanno comunicato il peso tre volte 404 (21%).

    Questo gruppo intermedio si suddivide ulteriormente in base **alla regione di appartenenza** nei seguenti due gruppi che sono anche nodi terminali (T10,T11):

    - **nodo terminale T10** quelli residenti in: Abruzzo, Basilicata, Calabria, Emilia Romagna, Lombardia, Piemonte, Puglia, Sardegna, Sicilia, Toscana, Trentino alto Adige;
    - **nodo terminale T11** quelli residenti in: Veneto, Friuli, Liguria, Marche, Umbria, Lazio, Campania.

Sintesi della segmentazione ad albero in funzione delle variabili caratteristiche

gruppo	numerosità	Prevalenza risposte al peso e loro incidenza nel gruppo	Peso del gruppo	!a Variabile caratteristica	2a Variabile caratteristica	3a Variabile caratteristica
T4	469	2 (44%)	9%	Life	Età >50 anni	-
T10	558	2 (37%)	10%	Life	Età <50 anni	Regione 1(*)
T11	516	3 (43%)	10%	Life	Età <50 anni	Regione 2
T3	3824	3 (56%)	71%	Platinum	-	-

Nota: Per età è stata considerata l'età all'attivazione del mantenimento (AM)
(*)Regione 1 : Abruzzo, Basilicata, Calabria, Emilia Romagna, Lombardia, Piemonte, Puglia, Sardegna, Sicilia, Toscana, Trentino alto Adige.
(**)Regione 2: Veneto, Friuli, Liguria, Marche, Umbria, Lazio, Campania.

Descrizione dei gruppi
1 gruppo nodo terminale (T4) di 469 soggetti (9%) del campione caratterizzato da una prevalenza di due risposte al peso, con tipo di contratto Life e con età minore di 50 anni;
2 gruppo nodo terminale (T10) di 558 unità pari al 10% del campione, caratterizzato da una prevalenza di due risposte al peso, età maggiore di 50 anni, residente in Basilicata, Campania, Emilia e Romagna, Lombardia, Piemonte, Marche, Toscana;
3 gruppo nodo terminale (T11) di 516 unità pari al 10% del campione, caratterizzato da una prevalenza di tre risposte al peso, età maggiore di 50 anni, NON residente in Basilicata, Campania, Emilia e Romagna, Lombardia, Piemonte, Marche, Toscana;
4 gruppo nodo terminale (T3) di 3824 unità pari al 75%% del campione, caratterizzato da una prevalenza di tre risposte con contratto Platinum

In sintesi, sono più "ligi", nel comunicare il proprio peso, i clienti con contratto *platinum* se confrontati con i clienti *life*. Nell'ambito dei clienti life (i più giovani) vi è una influenza della regione di residenza.
In conclusione le variabili che intervengono nel processo di bipartizione dei clienti sono, in ordine di importanza:

- **tipo di contratto:** (Platinum -71%: maggiormente disposti a fornire il peso e Live (29%) con comportamenti diversificati in funzione dell'età di ingresso nella fase di Mantenimento;
- **Age at AM** (age < di 50 anni, i più giovani, AM e age > di 50 anni più anziani);
- **Regione di residenza** che interviene nel 20% del campione

Considerando le due analisi si può concludere che il fenomeno delle mancate risposte al peso è connesso con diverse gradi di importanza alle seguenti variabili: Durata in FI, Tipo di contatto, età

di ingresso nella fase di mantenimento e regione di residenza. In particolare va considerato il tipo di contratto e l'età di entrata in Mantenimento come variabili warning.

5 -Modello logistico

Constatato che è presente, nel set di dati fornito da Bioimis, circa il 53% di 'mancate risposte', si ritiene utile, ai fini di una conoscenza più dettagliata di questa tipologia di cliente, di utilizzare una metodologia statistica che individui le caratteristiche dei soggetti propensi a fornire *missing* con associata una probabilità. I tal modo i gestori del data base potranno mettere in atto azioni rivolte al miglioramento di questa lacuna e definire regole per ottenere una base dati più informativa e corretta ai fini statistici.

La regressione logistica: alcuni concetti generali

-*La regressione logistica* fa parte dell'insieme dei modelli lineari utilizzati in medicina è del tipo:

$$\text{Log}\,[\pi/(1-\pi)] = B_0 + B_1 X_1 \ldots B_i X_i \ldots B_k X_k + \varepsilon$$

Dove :

a) $\boldsymbol{\pi}$ è la probabilità che la variabile risposta assuma valore 1 *(ci sia un dato mancante)*;

b) le $\mathbf{X_1 \ldots X_i \ldots X_k}$ sono le k variabili esplicative *(ossia che spiegano la variabile m.r.)*;

c) ε è la parte del modello non spiegata dalla variabili esplicative, generalmente denominata 'caso'.

d) $\mathbf{B_i}$ sono i coefficienti che indicano l'effetto che ogni variabile ha sulla variabile risposta *(possono essere positivi o negativi)*.

La variabile risposta associata alla probabilità π può assumere due valori (0;1) ossia assenza/ presenza del requisito indicabile con Y(m.r.). Il valore di π esprime la probabilità che si verifichi l'evento Y (ossia presenza di m.r.), (1-π) la probabilità che non si verifichi l'evento Y.

La quantità Log (π/1-π) è definita come log-odds, dove 'odds' dell'evento Y ha il seguente significato: è il rapporto tra la probabilità π di un evento e la probabilità che tale evento non accada (1-π), ossia il suo complementare. Il valore del rapporto può essere maggiore, uguale o minore di 1 ed in tal senso acquisisce il suo significato: se il valore è maggiore di 1 significa che l'evento Y ha maggiore probabilità di verificarsi rispetto a non verificarsi. Nel nostro caso è più probabile avere una mancata risposta rispetto ad avere la risposta.

Nel caso in esame la variabile Y è costituita dalla '*mancata risposta'* al peso a 60gg e 90gg che costituisce un elemento determinate sulla qualità della base dati esaminata. E' importante per Biomis conoscere quali siano le variabili esplicative che maggiormente influiscono sulle mancate risposte, sopra accennate, al fine di aumentare la qualità dei propri dati.

A puro scopo esemplificativo si supponga di voler studiare l'effetto di una cura su un insieme di pazienti trattati con un farmaco e pazienti trattati con un placebo. La probabilità sopra accennata riflette il seguente concetto: supponiamo che tra i pazienti sottoposti a cura l'80% abbia una remissione dei sintomi. In questo caso **l'odds** per i soggetti sottoposti a cura è 0,8/(1-0,8) = 4 che significa: la probabilità di remissione dei sintomi per i pazienti trattati è di 4 a 1 rispetto ai non trattati. *Volendo fare un esempio relativo alle corse dei cavalli: due cavalli sono quotati A) 4 a 1 e B) 20 a 1. Ciò significa che il cavallo nella situazione A ha maggiore probabilità di vincere rispetto*

al cavallo nella situazione B o inversamente il cavallo della situazione B ha un probabilità di non vincere 20 volte superiore a quella di vincere. Si conclude che il cavallo più forte è quello della situazione A caratterizzato da minore probabilità di non vincere e di riflesso il suo complemento è la maggiore probabilità di vincere. Ovviamente un giocatore puntando 1 euro sul cavallo della situazione A ne vince solo 4 mentre ne vince 20 se punta sul cavallo della situazione B e questo vince.

Riprendendo il concetto di regressione logistica, essa studia la relazione di dipendenza del possesso di un attributo dicotomico (nel nostro caso presenza/assenza di un dato mancante), in funzione di una o più variabili indipendenti ($X_1 \ldots X_i \ldots X_k$) di natura qualsiasi (qualitative e/o quantitative). Nel caso dei dati Bioimis si tratta di studiare le mancate risposte fornite dai clienti a vari momenti del "programma alimentare". Si valuterà l'effetto delle variabili indipendenti considerate, sulla probabilità di non fornire una risposta rispetto a quella di fornire il dato richiesto. Lo scopo è di individuare tra le variabili indipendenti, quelle che hanno maggiore potere esplicativo sulla variabile risposta, ossia quelle che vanno interpretate come determinanti del possesso o meno del requisito.

La relazione può essere definita da una correlazione positiva o negativa con la variabile dipendente e possono essere considerate come fattori di rischio o di protezione. In conclusione si va a cercare quale sia la migliore combinazione lineare delle variabili indipendenti che meglio discrimina le unità statistiche tra quelle che possiedono l'attributo e quelle che non lo possiedono.

Il modello stima le probabilità del possesso dell'attributo ***(in questo caso di non dare una risposta***) per una nuova unità statistica di cui è stato osservato l'insieme dei dati indipendenti e fissato per tale probabilità un valore soglia. Ossia permette di classificare una nuova US nella categoria delle US che possiedono l'attributo o in quelle che non lo possiedono.

-*Bontà di adattamento del modello* è valutata con diversi metodi, uno dei più usati e il metodo AIC (Akaike Information Criterium) (cfr. H. Akaike, 1974). Questo criterio fornisce una misura della qualità della stima di un modello statistico tenendo conto sia della bontà di adattamento che della complessità del modello. E' infatti noto che all'aumentare delle variabili esplicative aumenta la bontà di adattamento del modello teorico ai dati osservati. Va cercato il giusto compromesso tra l'ottenere un buon adattamento, tra i dati osservati e i dati predetti senza incorrere nel problema dell'overfitting, cioè nel cercare di spiegare quella parte dei dati che è dovuta puramente al caso e non è quindi sistematizzabile. A tal fine gli indici di adattamento basati sulla verosimiglianza penalizzano la verosimiglianza del modello con una parte che dipende dalla complessità del modello, solitamente funzione del numero di variabili utilizzate e del numero di unità statistiche considerato. Poiché tali indici usano sistemi di penalità diversi il loro significato ha senso solo in un'ottica comparativa e non assoluta.–

-*Criteri di selezione delle variabili (forward, backward, stepwise).*Il metodo **forward** (in avanti) inizia con un modello 'nullo' nel quale nessuna variabile è selezionata tra i predittori. Nel primo step viene aggiunta la variabile con l'associazione maggiormente significativa sul piano statistico. Ad ogni step successivo è aggiunta la variabile con la maggiore associazione statisticamente significativa tra quelle non ancora incluse nel modello, ed il processo prosegue sino a quando non vi sono più variabili con associazione statisticamente significativa con la variabile dipendente. Il metodo **backward** (all'indietro) inizia con un modello che comprende tutte le variabili e procede, step by step, ad eliminare le variabili partendo da quella con l'associazione, con la variabile dipendente, meno significativa sul piano statistico. Il processo **stepwise** fa avanti e indietro tra i due processi, aggiungendo e rimuovendo le variabili che, nei vari aggiustamenti del modello (con

aggiunta o re-inserimento di una variabile) guadagnano o perdono in termini di significatività. La regressione stepwise è un metodo di selezione delle variabili indipendenti allo scopo di selezionare un set di predittori che abbia la migliore relazione con la variabile dipendente. Nella presente applicazione si è utilizzato il processo stepwise.
Al termine del processo il sistema produrrà un output con: a) le variabili indipendenti (covariate) che spiegano meglio il modello, b) il loro potere esplicativo (coefficienti) inteso come la forza dell'effetto sulla variabile risposta dovuta alla variabile esplicativa X_i, c) il livello di significatività (probabilità che quell'effetto sia solo dovuto al caso).

L'applicazione del modello logistico ai dati Bioimis: risultati

Il modello logistico è stato applicato al campione di dati (5367 casi) estratto dal DB Bioimis. In questo campione sono presenti mancate risposte del peso: all'inizio della FI, a 30gg, a 60gg ed a 90 gg.. Vista la maggiore frequenza di m.r. a 60gg e 90gg si sono stimati due modelli uno applicato alle risposte del peso a 60gg e l'altro a 90gg : In pratica si sono utilizzati due files differenti derivanti dalla selezione dal file iniziale di 5367 casi.
La variabile risposta è il 'missing del peso al tempo 't' (t=60gg o 90gg) mentre le variabili esplicative, scelte per la stima dei parametri, sono state individuate avendo avuto cura di eliminare:

a) le variabili altamente correlate tra di loro (age at AM e age at UP che presentavano un'alta correlazione con age at IFI),
b) la "durata FI" che presenta troppi missing con conseguente influenza sul calcolo dei parametri.

Pertanto le variabili esplicative in analisi sono state le seguenti:

1. Stato di salute
2. Regione residenza
3. Statura
4. Age at IFI
5. Customers Sex
6. Polso
7. BMI Iniziale
8. Tipo Contratto

Il criterio stewise utilizzato procede nel seguente modo: considera inizialmente tutte le variabili ed elimina, passo dopo passo, le variabili con minor perso nella esplicazione della variabile dipendente. Il che significa che elimina secondo il criterio AIC la variabile con contributo di devianza minore. Il criterio non elimina definitivamente la variabile con devianza minore ma può, ai passi successivi, reinserirla per migliorare il 'fitting' del modello. Quindi in parole povere toglie e mette da parte le variabili e confronta con altre soluzioni cercando la configurazione migliore, delle covariate, che meglio spiega la variabilità del modello. Il software utilizzato fornisce il modello migliore sulla base dell'AIC.

Di seguito sono riportati i principali risultati dei due modelli finali considerando i missing del peso a 60gg ed a 90gg.

A) ***Modello a 60gg*****.**

Le variabili statura, regione di residenza, stato di salute sono state eliminate per la loro scarsa importanza in base al criterio AIC, pertanto il modello finale risultata il seguente:

Call: glm(formula = nmis60 ~ Tipo.Contratto + BMI_Iniziale + Polso + Age.at.IFI, family = "binomial", data = an)

Tabella 1

Coefficients				
	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	-1.599939	0.465822	-3.435	0.000593***
PLATINUM	-0.925710	0.110550	-8.374	< 2e-16***
BMI iniziale	-0.028807	0.009573	-3.009	0.002620**
Polso	0.064200	0.031955	2.009	0.044530*
Age at IFI	0.007320	0.003234	2.263	0.023624*

Signif. codes: 0 '***' ; 0.001 '**' ; 0.01 '*'; 0.05 '.' ; 0.1 ' ' 1

Number of Fisher Scoring iterations: 4

Tabella 2

Exp (Coef (mod60s))- **Odds**				
Intercept	Contratto PLATINUM	BMI iniziale	Polso	Age at IFI
0.2019088	0.3962499	0.9716039	1.0663056	1.0073464

La tabella 1 riassume tutti gli elementi per individuare le variabili più significative ai fini di individuare quali siano le variabili più rilevanti collegati ad una mancata risposta a 60gg.
In particolare: a) *estimate* rappresenta il log-odds dell'effetto che la variabile esplicativa riportata a sinistra nella prima colonna ha sulla variabile risposta. Nel nostro caso il tipo di contratto PLATINUM riduce la probabilità di fornire un dato mancante a 60gg (in letteratura viene considerato come fattore protettivo); b) *Standard error* è la variabilità della distribuzione dello stimatore di B_i (Platinum, BMI, Polso, Age at IFI)); c) *z value* è il corrispondente valore standardizzato utilizzato dal programma per calcolare il *p.value*: Pr.(>|z|); d) gli asterischi indicano il grado di significatività ossia la probabilità di ottenere un valore di "z value" maggiore di quello trovato. Maggiore è il numero degli asterischi è migliore è la significatività della stima ottenuta.
Conclusioni emergenti dal modello costruito sulla probabilità di missing del peso a 60gg:
1-Il tipo di programma **platinum** presenta una probabilità' più bassa rispetto al' life' di fornire mancate risposte a 60gg. Più in dettaglio, a parità di tutte le altre condizioni il rapporto degli odds tra chi è ***platinum*** e chi è ***live*** di fornire un dato mancante a 60gg è di circa 0,4 (cfr.tab.2). Quindi chi appartiene al programma platinum ha una probabilità di 0.4 volte superiore di fornire un dato

mancante rispetto a chi appartiene al programma live, o, analogamente, chi appartiene al programma live ha una probabilità di fornire un dato mancante a 60 gg rispetto al non fornirlo di 2,5 volte (1/0,4=2,5) superiore a chi appartiene al programma platinum.
2- In linea di principio ci si aspetterebbe che, più una persona è in sovrappeso e più sia ligia nel partecipare al programma e fornire i dati richiesti e quindi la probabilità di dare informazioni, rispetto al non fornirle, dovrebbe aumentare. Infatti, il coefficiente connesso al BMI è negativo e il corrispondente odds è minore di uno e quindi la variabile "BMI iniziale" ha un effetto negativo sulla probabilità di dare mancate risposte a 60gg, quindi è più probabile che una persona con alto BMI iniziale fornisce le informazioni richiesta a 60gg. Ciò significa che per ogni incremento unitario del **BMI** iniziale, a parità di tutte le altre condizioni, la probabilità di fornire un dato mancante a 60gg, rispetto al non fornirlo, si riduce di circa il 3% (che è 1-0,97).
3- Un'altra variabile da considerare è **l'età a IFI** *(età inizio forma ideale),* sostanzialmente gli anziani sono leggermente più propensi a non fornire dati a 60gg a parità di tutte le altre variabili.
4- Va osservato anche l'effetto significativo della misura del polso, che a parità di tutte le altre condizioni, provoca, per ogni aumento unitario della circonferenza, un aumento del 7% delle probabilità di fornire un missing rispetto a non fornirlo a 60gg. Quanto sopra descritto dal modello, per questa variabile non trova a nostro avviso una giustificazione logica e probabilmente può essere dovuto ad un effetto di mascheramento della variabile genere dato che mediamente gli uomini hanno un polso più grande delle donne.

B) Modello a 90 gg

Le variabili statura, regione di residenza, stato di salute sono state eliminate per la loro scarsa importanza in base al criterio AIC, pertanto il modello finale risultata il seguente

Call: glm(formula = nmis90 ~ Tipo.Contratto + BMI_Iniziale + Polso + Age.at.IFI, family = "binomial", data = an2)

Tabella 3

Coefficients				
	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	-0.708488	0.338345	-2.094	0.03626
PLATINUM	-0.857123	0.083150	-10.308	< 2e-16
BMI iniziale	-0.017077	0.006283	-2.718	0.00657
Polso	0.050926	0.023224	2.193	0.02832
Age at IFI	0.013747	0.002398	5.733	9.88e-09

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of Fisher Scoring iterations: 4

Tabella 4

Exp(Coef(mod90s)) **Odds**				
Intercept	Contratto PLATINUM	BMI iniziale	Polso	Age at IFI
0.4923881	0.4243813	0.9830681	1.0522452	1.0138422

Conclusioni emergenti dal modello costruito sulla probabilità di missing del peso a 90gg:
Dai logodds , si evince che il comportamento dei clienti a 90 gg ricalca quello a 60gg anche se con intensità leggermente diverse. Si riduce di poco il divario tra i tipi di contratto, si riduce l'effetto del BMI ed aumenta, anche se di poco, l'effetto dell'età sulla probabilità di fornire un dato mancante. Nel lungo periodo è più facile che un anziano non fornisca risposte.

Considerazione generale:
I risultati del modello logistico non andrebbero interpretati in termini assoluti confrontando i coefficienti delle varie covariate direttamente tra di loro ma con riferimento alla tipologia di variabile considerata che risulta contraddistinta da unità di misura e campo di variazione. Ad es: l'effetto di un aumento del 5% della probabilità di fornire un missing per un aumento unitario della misura della circonferenza del polso a 90 giorni *potrebbe sembrare più forte* di un aumento dell'1% della stessa probabilità di fornire un dato mancante (rispetto al non fornirlo) per un aumento di un anno dell'età. Va osservato che le età (espresse in anni compiuti) hanno probabilmente un campo di variazione più ampio della circonferenza del polso (espresse in cm.), per cui è più facile trovare clienti con differenze di età di 10 anni (che producono un effetto sulla variazione dell'odds di fornire un dato mancante a 90 giorni di circa il 14%) rispetto a trovare persone con circonferenze del polso con una variazione di circa 3 cm che è quello che a grosse linee fornirebbe lo stesso effetto sulla variazione dell'odds di fornire un dato mancante a 90 giorni.
Separatamente, in allegato viene fornita l'applicazione per calcolare la probabilità di fornire una mancata risposta in base ai valori delle seguenti variabili:

- Tipo di contratto
- BMI iniziale
- Polso
- Età ad IFI

6 - La Cluster Analysis

Al fine di affinare l'analisi sui clienti del campione Bioimis, si è proceduto alla applicazione di una ulteriore tecnica statistica sui soli clienti che hanno fornito il loro peso alle date prestabilite dal loro programma alimentare. Si è scelto in un primo momento una metodologia di classificazione basta sul criterio della individuazione di particolari gruppi di clienti non considerando il peso come variabile dipendente da altre covariate ma inserendolo direttamente nel set delle variabili senza considerare una relazione tra esse. A tal fine si è utilizzata la cluster Analysis.

Generalità

La cluster analysis è un metodo statistico esplorativo multivariato che si prefigge di raggruppare le unità statistiche, in modo da minimizzare la "lontananza logica" interna in ciascun gruppo e di massimizzare quella tra i gruppi. La "lontananza logica" viene quantificata per mezzo di misure di similarità/dissimilarità definite tra le unità statistiche (clienti). Lo scopo è quello di ricercare l'esistenza di gruppi di individui non noti a priori o non identificabili direttamente, sulla base di una serie di variabili (quantitatve/qualitative) che a volte sono denominati 'indicatori'. L'indicatore può essere ottenuto anche da trasformazioni delle variabili di partenza *(cfr. Cluster Analysis- B.S. Everitt-Hodder Arnold ed 1993; Cluster Analysis- 5th ed.- B.S. Everitt ed altri- Wiley 2011).*

Gli elementi essenziali di un processo di cluster sono la scelta:

a) delle osservazioni da sottoporre ad analisi (campione / popolazione),

b) degli 'indicatori' (variabili quantitative/qualitative o loro elaborazione),

c) delle misure di similatità/dissimilarità (metriche),

d) dell'algoritmo che permette di effettuare i raggruppamenti (criterio do costruzione dei gruppi),

e) del numero dei gruppi,

f) della valutazione dei risultati (in genere connessa alla metrica).

Gli algoritmi sono classificabili in due categorie:

a) gerarchici : permettono di costruire una gerarchia (albero/dendrogramma) tra i gruppi secondo un sistema aggregativo (botton-up) o scissorio (top-dwon) questi ultimi richiedono molto tempo di calcolo e quindi sono i meno utilizzati. Forniscono una "famiglia" di partizioni partendo da quella banale in cui tutti gli elementi sono distinti (i "gruppi" coincidono con le unità/soggetti) sino a quella in cui tutte unità sono riunite in un unico gruppo. Nel primo passo si uniscono le unità con minore distanza che formano il primo gruppo. I gruppi che si ottengono ad un certo stadio possono solo essere riuniti nei passi successivi (ma non scissi). Tutti i metodi gerarchici permettono di scegliere a 'posteriori' il numero di gruppi più significativo e sono costruiti in modo tale che la partizione in K-1 gruppi (P,k-1) contenga la partizione in K-gruppi (P,k*) es. un taglio dell'albero in 2 gruppi contiene anche il taglio in 3 gruppi.* Gli algoritmi gerarchici sono i più utilizzati per la loro facilità di calcolo e via via sono state proposte nuove metriche al fine di rendere più significative le 'partizioni' ossia: i soggetti appartenenti ad un gruppo (A) sono tra loro più simili (secondo la metrica utilizzata) rispetto al confronto tra ciascuno di essi e un qualsiasi altro soggetto appartenente ad un *gruppo diverso da (A). Questa particolarità è la caratteristica di una partizione ben strutturata. Ossia si ha la massima omogeneità interna a la massima disomogeneità tra gruppi.*

b) non gerarchici : forniscono direttamente il numero dei gruppi che è scelto a priori. Generalmente questi metodi non sono utilizzati se non in particolari casi particolari.

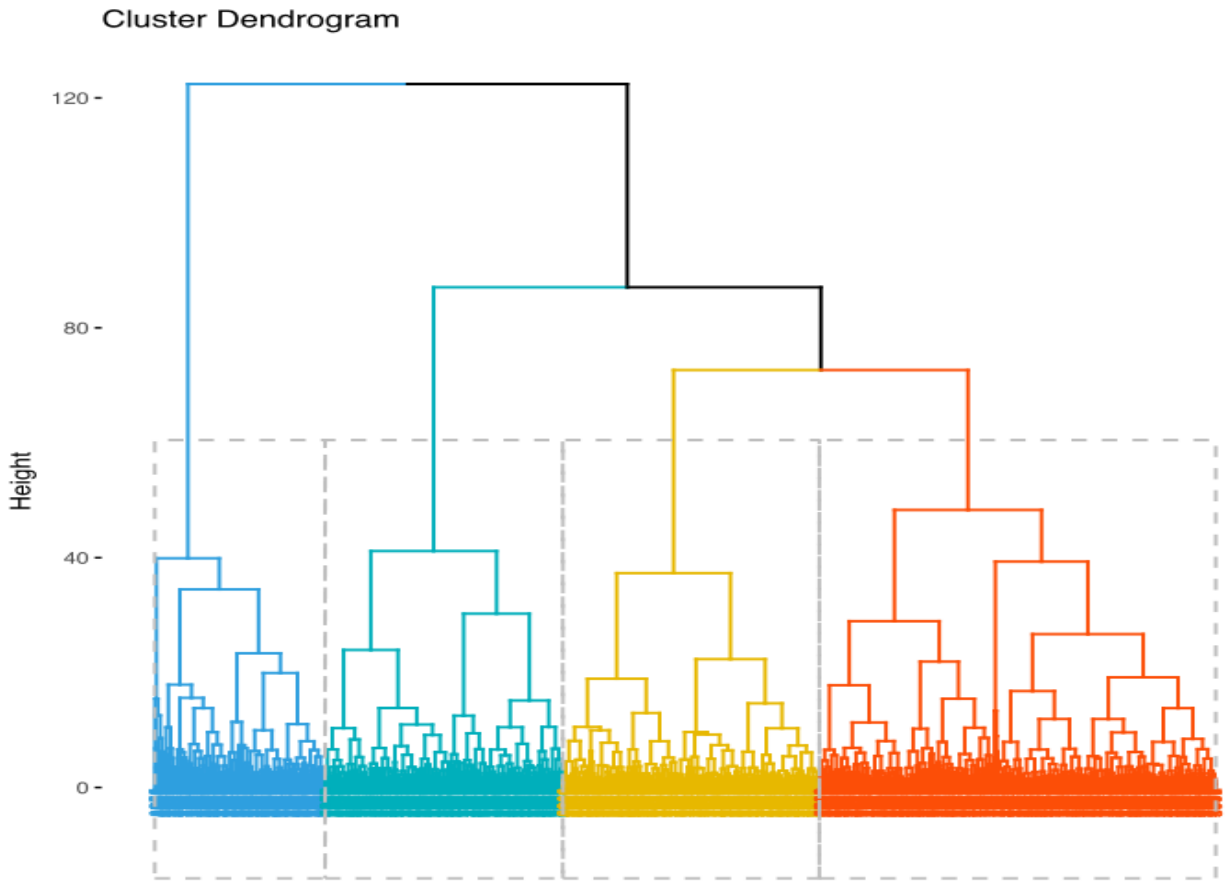
I risultati del processo di cluster sono frutto delle scelte degli elementi essenziali.

Per questa analisi si è scelto il criterio di Ward (*introdotto nel 1963 cfr. Cluster Analysis- 5th ed.- B.S. Everitt ed altri- Wiley 2011*) che risulta essere il più efficace tra i metodi aggregativi e fornisce partizioni ben strutturate. L'algoritmo parte dagli N individui costituenti l'intero campione e li aggrega via via, uno alla volta. Nel primo passo si sceglie la partizione in N-1 gruppi, che minimizzi sostanzialmente la somma delle varianze dei due gruppi che si sono formati. Quindi, generalizzando, si passa da una partizione K (dove k è il numero dei gruppi) la migliore partizione in k-1 gruppi cercando di ottimizzare la funzione obiettivo, definita come differenza tra le varianze/ devianze delle partizioni in K gruppi e k-1 gruppi. Con questo criterio si hanno due vantaggi: a) i gruppi che presentano varianza minima all'interno di essi quindi sono i più omogenei ossia i

soggetti appartenenti al gruppo presentano un vettore di informazioni molto simile al vettore medio; b) è massima la diversità tra i gruppi che significa che vi sono elementi distintivi da gruppo e gruppo. Il processo si ferma finché non si sono raggruppate tutta le unità in un solo gruppo.
L'output è costituito in generale da un grafico 'dendrogramma' che, da un lato riporta le unità del campione/popolazione e dall'altro il livello di aggregazione delle stesse (metrica). Scegliendo un determinato livello di aggregazione si ottiene il taglio dell'albero e vengono così individuati il numero di gruppi ottimali per il livello scelto. La forma dell'albero indica generalmente il numero ottimale di gruppi.

La cluster sul campione di dati Bioimis

Il metodo di Ward è stato applicato all'insieme degli individui, composto da 2684 clienti che hanno fornito 4 risposte alla richiesta del loro peso: iniziale, a 60gg., a 90gg., ultimo peso. Questo campione costituisce circa il 47% dell'intero insieme dei dati 'anagrafica' fornita da 'Bioimis'.
L'insieme delle variabili considerate per la ricerca dei gruppi è così costituito:
a) sesso, stato di salute, regione di residenza, (ripartizione geografica), tipo di contratto, età in anni compiuti *(considerate come variabili descrittive/supplementari);*
b) BMI iniziale, peso iniziale, peso a 60gg, peso a 90gg, ultimo peso, polso, altezza, età entrata IFI, età entrata AM, età entrata UP, durata FI *(considerate come variabili in analisi)*.
Le variabili in analisi sono state standardizzate al fine di eliminare le diverse unità di misura.
L'output ha prodotto la sequenza di partizioni nel grafico sotto riportato dal quale emerge che la migliore partizione è costituita da 4 gruppi che vengono di seguito descritti e riassunti in una successiva tabella. La partizione ottenuta evidenzia un ordinamento dei gruppi (esclusa la denominazione che non è automatica) è possibile quindi ordinarli in modo crescente rispetto ai valori delle variabili : BMI iniziale, peso iniziale, peso a 66gg, peso a 90gg, ultimo peso, polso, altezza.



asse degli individui classificati in gruppi

Descrizione dei 4 gruppi individuati

Il gruppo 3 *(evidenziato nel grafico in colore giallo scuro)* : composto da 649 individui (24% del campione) presenta valori medi, delle variabili considerate e indici di variabilità per :BMI iniziale, peso iniziale, peso a 60gg, peso a 90gg, ultimo peso, polso, altezza, *inferiori* agli altri gruppi. I soggetti di questo gruppo presentano *un'età media elevata* 48 anni rispetto agli altri gruppi per le variabili: IFI,AM,UP e presentano un *più breve periodo nella forma ideale d*i 57,6 giorni prima di entrare nella fase di mantenimento.
Praticamente sono i soggetti che iniziano il trattamento non avendo grossi problemi di peso e valori di BMI appena al di spora del limite consentito dalla letteratura.
Il gruppo è così composto:

a) prevalentemente da Femmine (95%) ben al di sopra dell'80% (valore medio del campione);
b) Prevalente da contratti LIFE 57% (al di sopra del 25% del campione);
c) Con residenza al NE 26% (prevalente rispetto alla media del 2% del campione);
d) Presenza di Ipo tiroidei (27% di tutti gli ipotiroidei)

Gli appartenenti a questo gruppo si possono denominare: **appena in sovrappeso**

Il gruppo 4 *(evidenziato nel grafico in colore celeste scuro)*: composto da 431 individui (16% del campione) presenta valori medi delle variabili considerate e indici di variabilità per: BMI iniziale, peso iniziale, peso a 60gg, peso a 90gg, ultimo peso, polso, altezza, *superiori* agli altri gruppi. I soggetti di questo gruppo presentano una età media di 42 anni per le variabili : IFI,AM,UP e **un più elevato** tempo nella Forma Ideale 81,6 giorni prima di entrare nella fase di mantenimento.
Praticamente sono i soggetti che iniziano il trattamento avendo grossi problemi di peso e valori di BMI molto al di spora del limite consentito dalla letteratura.
Il gruppo è così composto:

a) prevalentemente da maschi (50%) ben al di sopra del 20% valore medio del campione;
b) solo contratti PLATINUM ;
c) con residenza al SUD (50%) e nelle isole (14%) sempre rispetto alla media del campione;
d) I sani sono sotto rappresentatati 14% rispetto al 66% del campione

Gli appartenenti a questo gruppo si possono denominare **: in forte sovrappeso**

Il gruppi 2 e 1 presentano valori abbastanza simili per le variabili esaminate ma quasi sempre il secondo gruppo evidenzia situazioni meno critiche. L'insieme dei due gruppi conta 1652 individui e per ciascuna variabile si presentano valori di variabilità abbastanza simili mentre i valori medi sono così rappresentati.:

Il gruppo 2 *(evidenziato nel grafico in colore rosso)* : composto da 1003 individui (37% del campione) è il più numeroso e presenta valori medi delle variabili considerate e indici di variabilità per: BMI iniziale, peso iniziale, peso a 60gg, peso a 90gg, ultimo peso, polso, altezza, *superiori* solo al terzo gruppo. I soggetti di questo gruppo presentano *l'età media più elevata* di 51 anni per le variabili :IFI,AM,UP AM e la più elevata permanenza nella forma ideale di 82,5 giorni prima di entrare nella fase di mantenimento.
Sono i soggetti che iniziano il trattamento avendo valori di BMI al di spora del limite consentito dalla letteratura.
Il gruppo è così composto:

a) Non si registra una prevalenza della tipologia di sesso rispetto al campione
b) Prevalente del contatto PLATINUM 95% (del totale platinum) (al di sopra del 80% del campione);
c) Con leggera prevalenza della residenza al NW e NE rispetto alla media del campione);
d) Presenza di Ipertesi (58% degli ipertesi)

Gli appartenenti a questo gruppo si possono denominare **:in sovrappeso accentuato**

Il gruppo 1 *(evidenziato nel grafico in colore verde* **)** : composto da 601 individui (22% del campione) presenta valori medi delle variabili considerate e indici di variabilità per : BMI iniziale, peso iniziale, peso a 60gg, peso a 90gg, ultimo peso, polso, altezza, inferiori rispetto ai gruppi 2 e 4. I soggetti di questo gruppo presentano l'età media più bassa 29 anni per le variabili: IFI, AM,UP e di un periodo nella forma ideale di 72,1 giorni prima di entrare nella fase di mantenimento è inferiore ai gruppi 2 e 4.
Praticamente sono i soggetti che iniziano il trattamento con valori di BMI al di spora del limite consentito dalla letteratura.
Il gruppo è così composto:

a) Non si registra una prevalenza della tipologia di sesso rispetto al campione;
b) Non c'è una prevalenza del tipo di contratto rispetto alla media del campione;

c) Prevalenza di residenti al Sud mentre una scarsa presenza dei residenti al centro;
Gli appartenenti a questo gruppo si possono denominare**: in sovrappeso**

Tabella riassuntiva dei gruppi con valori medi delle variabili

Variabili	Gruppo 1	Gruppo 2	Gruppo 3	Gruppo 4
BMI_Iniziale.mean	**32,4**	**34,8**	**27,8**	**44,3**
BMI_Iniziale.sd	4,7	4,6	3,5	6,9
Peso_Iniziale.mean	**91,3**	**93,9**	**71,6**	**130,8**
Peso_Iniziale.sd	13,7	11,1	7,9	20,3
Peso_A_60.mean	**81,2**	**83,5**	**64,2**	**115,9**
Peso_A_60.sd	11,7	9,9	6,9	18,4
Peso_A_90.mean	**79,5**	**81,3**	**63,3**	**112,5**
Peso_A_90.sd	11,1	9,8	6,7	18,1
Ultimo.Peso.mean	**79,9**	**82,0**	**64,6**	**110,6**
Ultimo.Peso.sd	12,1	10,5	7,4	17,9
Polso.mean	**17,1**	**17,7**	**16,2**	**19,4**
Polso.sd	1,2	1,3	1,0	1,6
Altezza.mean	**168,1**	**164,5**	**160,8**	**172,1**
Altezza.sd	7,9	7,0	6,7	9,0
Age.at.IFI.mean	**29,8**	**51,8**	**47,8**	**42,0**
Age.at.IFI.sd	8,7	9,3	9,0	10,5
Age.at.AM.mean	**29,9**	**51,9**	**48,0**	**42,2**
Age.at.AM.sd	8,7	9,2	9,1	10,5
Age.at.UP.mean	**30,7**	**52,9**	**49,0**	**43,1**
Age.at.UP.sd	8,8	9,2	9,1	10,5
DurataFI.mean	**72,1**	**82,5**	**57,6**	**81,6**
DurataFI.sd	24,8	37,3	22,2	22,0

I risultati dell'applicazione del metodo di Ward si possono sintetizzare nella seguente tabella:

Gruppi	**Variabili A)** ***valori medi***	**Variabili B)** ***età medie***	**Variabili C)** ***Durata FI (gg)***

Appena in sovrappeso *(gruppo 3)*	< degli altri gruppi	48 anni (elevata)	57,6 (la più breve)
Forte sovrappeso *(gruppo 4)*	>degli altri gruppi	42,5 anni	81,6 (molto elevato)
Accentuato sovrappeso *(gruppo 2)*	Medio alti rispetto altri gruppi	52 anni (+ elevata)	82 (+ elevato)
In sovrappeso *(gruppo 1)*	Medio bassi rispetto altri gruppi	29 anni (+ bassa)	71 (medio)

Nota:
Variabili A): BMI iniziale, peso iniziale, peso a 60gg, peso a 90gg, ultimo peso, polso, altezza.
Variabili B): età media per: IFI, AM, UP.
Variabili C): Durata FI

Conclusione: l'analisi ha prodotto una partizione ben strutturata in 4 gruppi che si differenziano bene in funzione delle tre tipologie di variabili utilizzate: a), b), c). Infatti, emerge un chiaro ordinamento in funzione del peso iniziale:

- *Appena in sovrappeso* il 25% del campione (G3) è composto da individui che presentano '*i più bassi valori'* delle variabili tipo a) rispetto agli altri gruppi, una minore permanenza in FI ed un'età media di 48 anni per le variabili tipo b). Con prevalenza di: Femmine, contratto LIFE, residenti nel NE, Ipo-tiroidei.
- *In sovrappeso* il 22% del campione (G1) è composto da individui che presentano '*valori medio bassi'* per le variabili di tipo a), una media permanenza in FI e la più bassa età media per le variabili di tipo b). Senza prevalenza di nessuna tipologia di sesso e di contratto, residenti al Sud.
- *Sovrappeso accentuato* Il 37% del campione (G2) è composto da individui che presentano '*valori medio alti '* per le variabili di tipo a), la più elevata permanenza in FI e la più elevata età media per le variabili di tipo b). Senza prevalenza di nessuna tipologia di sesso, ma prevalenza di contratto PLATINUM (95%), residenti al Nord ed Ipertesi.
- *Forte sovrappeso* il 16% del campione (G4) è composto da individui che presentano '*i più alti valori'* delle variabili di tipo a) rispetto agli altri gruppi, una elevata permanenza in FI ed con un'età media di circa 42 anni per le variabili tipo b). Con prevalenza di Maschi, residenti al Sud ed alle isole e senza patologie dichiarate**.**

Risultati in termini di perdita di peso medio sul campione esaminato

Considerando la differenza tra peso iniziale ed ultimo peso ci si deve statisticamente aspettare una variabilità dei risultati, dell'applicazione della dieta proposta, e quindi non tutti i clienti hanno presentato perdita di peso. Comunque le considerazioni vanno interpretate in termini di valori medi. L'analisi dei valori medi della perdita di peso nei quattro gruppi presenta la seguente situazione:

- il gruppo degli '**appena in sovrappeso'** presenta una perdita di peso di circa 7kg (circa 9% del peso iniziale);
- il gruppo dei '**sovrappeso**' presenta una perdita di peso di circa 11 Kg (circa 12% del peso iniziale);

- il gruppo dei' **sovrappeso accentuato**' presenta una perdita di peso di circa 11 Kg (circa 12% del peso iniziale
- il gruppo dei '**forti sovrappesi'** presenta una perdita di peso di circa 15 kg (circa 12% del peso iniziale);

Complessivamente utilizzando i valori medi i 2684 clienti hanno perso 28652 kg che corrispondono in media a 10,67 kg nel periodo medio di FI di 74 giorni.

7 - Analisi longitudinale

Con quest'ultima analisi si intende completare lo studio esplicativo della capacità della dieta, proposta dall'accademia alimentare Bioimis, a ridurre il peso dei soggetti che hanno perseguito la dieta proposta. Viene preso in esame l'andamento nel tempo della variabile oggetto di studio (peso) corredata da alcune variabili esplicative. A tal proposito si è considerata una particolare classe di modelli statistici detti 'modelli a misture finite' che consentono di gestire sia l'aspetto longitudinale dei dati sia di catturare l'eterogeneità del campione considerato in questa relazione. I modelli mistura costituiscono uno degli argomenti più interessanti ed informativi della modellistica statistica, come si può vedere da alcune interessanti monografie degli ultimi 20 anni [Titteringon ed al. (1985), McLachlan and Basford (1988), McLachlan and Peel (2000), Fruhwirth-Schnatter (2006)].
In particolare, il modello utilizzato considera una variabile aleatoria (in questo caso la variazione del peso) come frutto della 'mistura' tra più modelli statistici parametrici collegati alla stessa variabile osservata in diversi tempi di osservazione ed ad alcune variabili che si suppone possano spiegare le variazione della variabile risposta
Dato un insieme di dati $\mathbf{X} = (x_1, x_2, \ldots, x_n)$, dove *n* indica il numero delle osservazioni, si suppone che le istanze del data-set siano state prodotte a partire da una mistura di distribuzioni di probabilità. Lo studio si propone di individuare dei gruppi distinti tra di loro pensati come oggetti appartenenti ad una stessa tipologia di distribuzione, ma che si differenziano tra loro in base ai diversi valori assunti dai parametri che caratterizzano tale distribuzione (e.g. Gaussiana con parametri la media e la varianza) e aventi una loro probabilità di essere rappresentati.
In altre parole si cerca la tipologia di una funzione non nota f(x) della variabile aleatoria considerata *oggetto dello studio (variazione del peso*) che è frutto della combinazione di K distribuzioni note della stessa variabile ai vari tempi di osservazione.

L'obiettivo si concretizza nella ricerca della struttura di f(x), per definire successivamente dei gruppi, attraverso le osservazioni della stessa in più istanti di tempo, avendo come misura di approssimazione della validità del modello, l'adattamento della mistura ai dati osservati in base ad un indice legato alla diversa tipologia di mistura.

Nel caso di una mistura finita di distribuzioni continue, da noi scelto come elemento di analisi, la funzione di densità di probabilità cercata è descritta in generale da:

$$\mathrm{f}\left(\mathrm{x}, \vartheta_1 \ldots \vartheta_\mathrm{K}, \pi_1 \ldots \pi_\mathrm{K}\right) = \sum_{\mathrm{k}=1}^{\mathrm{K}} \pi_\mathrm{k} f_k\left(x, \vartheta_k\right)$$

dove (x) è la variabile considerata (variazione di peso), $\mathrm{f}\left(\mathrm{x}, \vartheta_1 \ldots \vartheta_\mathrm{K}, \pi_1 \ldots \pi_\mathrm{K}\right)$ è il modello statistico ignoto (che si vuole stimare attraverso le osservazioni del campione) con parametri $\vartheta_1 \ldots \vartheta_\mathrm{K}$, ciacuno con pesi $\pi_1 \ldots \pi_\mathrm{K}$. e componenti $f_k\left(x, \vartheta_k\right)$. Si sottolinea che, sebbene le $f_k\left(x, \vartheta_k\right)$ possano avere distribuzioni diverse, solitamente si sceglie di far assumere a tutte le $f_k\left(x, \vartheta_k\right)$ la stessa forma funzionale. I pesi $\pi_1 \ldots \pi_\mathrm{K}$ rappresentano il contributo individuale di ogni componente alla mistura finale. Ogni peso assume un valore non negativo compreso tra zero e 1 e tutti i pesi sono tali per cui

$$\sum_{\mathrm{k}=1}^{\mathrm{K}} \pi_\mathrm{k} = 1$$

I pesi $\pi_1 \ldots \pi_K$. prendono il nome di probabilità a priori e rappresentano la probabilità che un qualunque soggetto provenga da una qualunque delle componenti della mistura. Le componenti $f_k\left(x, \vartheta_k\right)$ si suppongono solitamente con distribuzione normale e parametri incogniti ϑ_k. Una volta scelta la forma funzionale per le $f_k\left(x, \vartheta_k\right)$ rimane il problema di determinare il numero ottimo di componenti K (definiti anche come 'gruppi') da utilizzare per approssimare al meglio la funzione f(x) cercata.

La scelta del numero di componenti avviene solitamente iterando da K=1 a un valore K_{max} (scelto a priori e comunque inferiore al numero di osservazioni *n*) e selezionando il numero di componenti che fornisce un valore ottimo della verosimiglianza penalizzata (tramite AIC *–Akaike's Information Crierium* o BIC- *Bayesian Information Criterium*). E' necessario fare ricorso alla verosimiglianza penalizzata perché è ovvio che, al crescere del numero delle componenti la mistura, ossia aumentando il numero dei parametri, aumenterebbe anche la verosimiglianza del modello con la conseguenza di una interpretazione dei dati complessa e poco informativa. Per tale motivo, in generale, si cerca un modello che dia il miglior valore di fitting *(adattamento dei dati al modello*) utilizzando un numero limitato di componenti (gruppi). Praticamente questo consiste nel cercare di determinare, all'interno del data set considerato, il minor numero di popolazioni con comportamento omogeneo delle unità al loro interno ma eterogenee tra di loro.

Poiché l'appartenenza ai gruppi è ignota, la stima dei parametri e il numero delle componenti viene realizzata in un'ottica iterativa utilizzano l'algoritmo EM (*Expectation-Maximization*) che si dimostra convergere a una stima localmente ottima dei parametri.

Come ulteriore output del modello, oltre alle probabilità a priori e ai parasmetri delle funzioni di densità di ogni gruppo, si ottengono le probabilità, a posteriori, per ogni singola unità, di appartenere a ciascuna delle componenti (gruppi o sottopopolazioni). Tali probabilità vengono indicate con il simbolo w_{ik} che rappresenta la probabilità che l'unità i-ma (cliente esaminato) appartenga alla componete (gruppo) k-ma della mistura. In simboli:

:

$$w_{ik} = \frac{\pi_k f_k(x_i)}{\sum_{j=1}^{K} \pi_j f_j(x_i)}$$

Nel nostro caso abbiamo scelto un K_{max} pari a 15 e abbiamo deciso di utilizzare l'indice BIC come indice di bontà di adattamento. Il modello che presenta il BIC più piccolo, tra tutti quelli considerati da K=1 a K=15, rappresenta il modello migliore. Nel nostro caso, utilizzando come variabile risposta la variazione relativa del peso, il criterio BIC ha suggerito un K uguale a 5, quindi indicando per la variabile peso una mistura a 5 componenti (che possono essere ugualmente denominati gruppi o sottopopolazioni).

Le variabili esplicative utilizzate per spiegare la variazione del peso nel tempo sono state:

- Sesso
- Tempo di osservazione
- Età alla forma ideale
- Stato di salute (la modalità "sano" è stata presa come modalità di riferimento)

I risultati ottenuti indicano che nei 5 gruppi così determinati l'effetto del tempo e delle altre covariate sulla variazione del peso differisce a seconda di quale gruppo venga considerato.

Come per ogni fenomeno statistico, esiste una variabilità e la variabile sottoposta ad esame non presenta comportamenti univoci in funzione delle covariate considerate. Pertanto delle 5 sottopopolazioni individuate 3 presentano gruppi di clienti che mostrano un effetto positivo nel tempo della dieta (gruppo, 1, 2, e 5), un gruppo non evidenzia effetti nè positivi nè negativi sulla efficacia della dieta, mentre un altro gruppo (gruppo 2) presenta un effetto negativo (il peso aumenta col passare del tempo).
Di seguito i risultati sintetici della elaborazione:

Risultati - variazione percentuali rispetto al basale

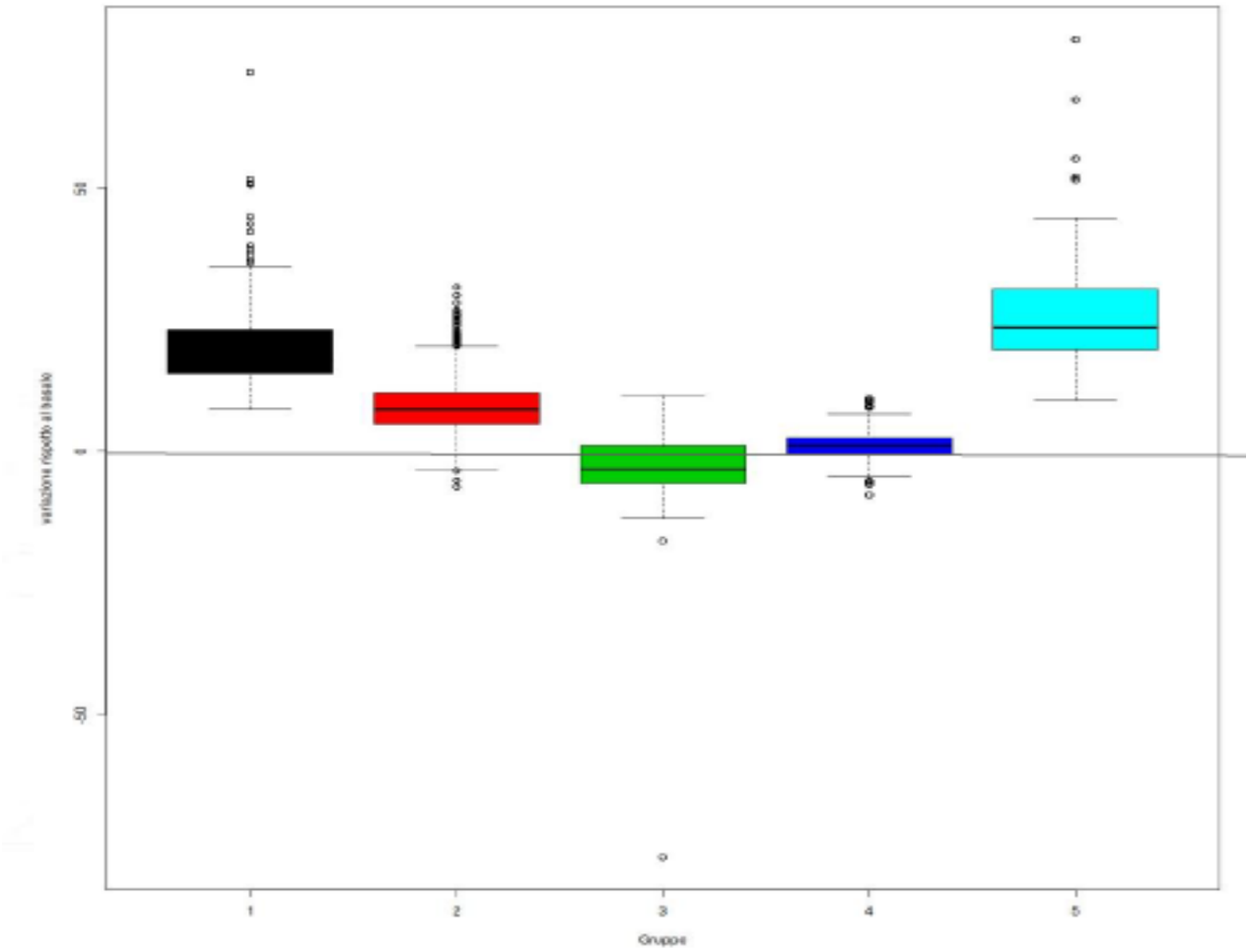
```
$Comp.1
      Estimate
(Intercept)                          -0.07384928
tempo                                -0.00030822
Customers.SexMaschio                 -0.00840811
Age.at.IFI                            0.0004114
I(Stato.di.salute == "Sano")TRUE      0.00563501
---

$Comp.2
      Estimate
(Intercept)                          -0.071376
tempo                                -0.000070844
Customers.SexMaschio                 -0.0081687
Age.at.IFI                            0.00016423
I(Stato.di.salute == "Sano")TRUE      0.0056289
---
```

```
$Comp.3
      Estimate
(Intercept)                          -0.051505
tempo                                 0.000068594
Customers.SexMaschio                 -0.010813
Age.at.IFI                           -0.0006094
I(Stato.di.salute == "Sano")TRUE      0.0099267

$Comp.4
      Estimate
(Intercept)                          -0.025834
tempo                                 0.000029554
Customers.SexMaschio                  0.00056099
Age.at.IFI                           -0.00027532
I(Stato.di.salute == "Sano")TRUE      0.0035738

$Comp.5
      Estimate
(Intercept)                          -0.0618336
tempo                                -0.00064964
Customers.SexMaschio                 -0.0167384
Age.at.IFI                            0.00036629
I(Stato.di.salute == "Sano")TRUE     -0.00147075
```

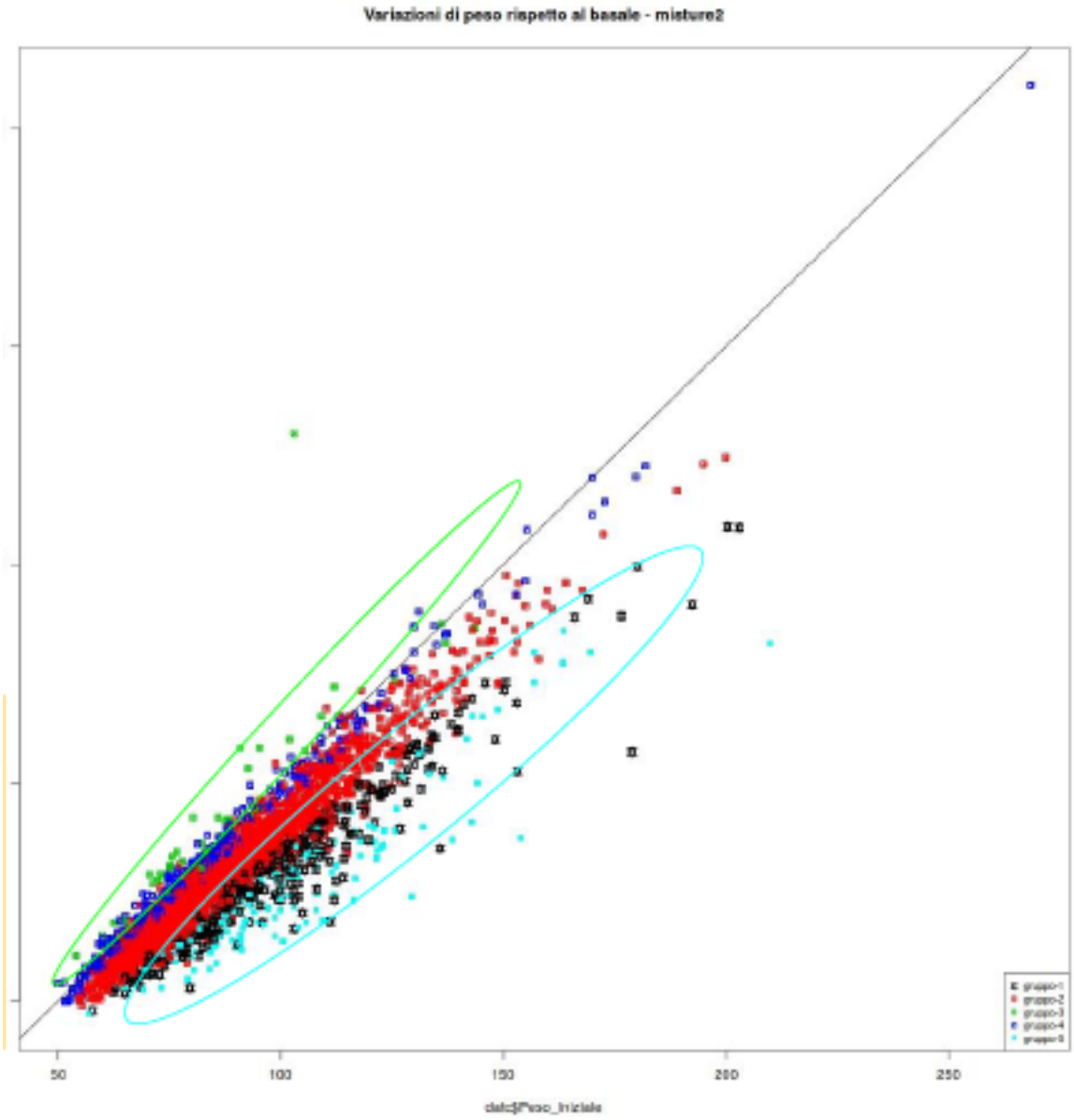
la descrizione dei gruppi individuati che ottimizzano il BIC viene riportata nel grafico seguente dove si evidenzia l'effetto della dieta con diminuzione di peso per i gruppi 1,2 e 5 evidenziati con il colore nero, rosso, celestino mentre il gruppo 4 i bleu è composto da soggetti che non hanno avuto grandi risultati dalla dieta. Il gruppo 5 mette in luce che la dieta seguita non ha prodotto significativi effetti.

Per quanto concerne la variabile tempo, un coefficiente negativo indica una riduzione del peso rispetto al peso iniziale che è tanto più alta quanto più alto in valore assoluto del coefficiente. Quindi il gruppo che presenta la maggiore riduzione di peso nel tempo è il gruppo 5, seguito immediatamente dal gruppo 1 e poi dal gruppo 2. I gruppi 3 e 4 presentano un effetto del tempo positivo, indicando, pertanto, un aumento del peso nel tempo rispetto al basale. Questo effetto è più pronunciato per il gruppo 3 e leggermente inferiore per il gruppo 4.
L'effetto del sesso è evidenziato dalla variabile dummy *(presenaza/assenza)* 'Customer.SexMaschio' che indica di quanto i maschi differiscano dalle femmine nella variazione di peso. Anche in questo caso un valore negativo indica che, in media, gli uomini hanno una variazione del peso, rispetto al basale, migliore rispetto a quella delle donne, cioè perdono più peso. Questo si verifica sostanzialmente per tutti i gruppi tranne che per il gruppo 4, in cui gli uomini si comportano peggio delle donne. Va detto che probabilmente gli uomini, partendo mediamente da un peso superiore rispetto alle donne, hanno una maggiore facilità a perdere più peso.
Stessa cosa per la variabile dummy 'Stato di Salute Sano' in cui si evidenzia l'effetto del non presentare alcune specifiche malattie (ipertensione, diabete etc.) sulla perdita di peso. In generale i "sani" perdono meno peso dei "malati", fatta eccezione per i pazienti del quinto gruppo.

La distribuzione della variazione del peso rispetto al basale per le 5 popolazioni individuate viene riportata nel grafico seguente dove si evidenzia l'effetto della dieta sulla diminuzione di peso per i gruppi 1,2 e 5 evidenziati con il colore nero, rosso, celestino mentre il gruppo 4, i bleu, è composto da soggetti che non hanno avuto grandi risultati dalla dieta. Il gruppo 3 mette in luce che la dieta seguita non ha prodotto significativi effetti. La variazione del peso rispetto al basale è stata rappresentata, per ogni gruppo individuato dal modello a mistura finita, utilizzando boxplot in cui sono stati riportati la mediana, il primo e terzo quartile della variazione del peso e i "baffi" rappresentano o il minimo e il massimo della variazione o il primo quartile meno 1.5 volte la differenza interquartilica (IQR) e il terzo quartile più 1.5 volte la differenza interquartilica. I punti fuori dai "baffi" rappresentano unità "anomale", cioè unità (clienti) che hanno valori troppo alti o troppo bassi rispetto a quelli delle altre unità della stessa distribuzione e che quindi dovrebbero essere considerate e studiate più in dettaglio.



Riportando in un diagramma cartesiano il peso iniziale (ascisse) e il peso finale (ordinate) di ogni individuo del collettivo considerato, i punti rappresentati hanno il seguente significato: gli individui sotto la bisettrice del primo quadrante rappresentano le persone per cui la dieta ha avuto successo. I punti intorno alla bisettrice, invece, sono individui per cui il peso all'inizio e alla fine della dieta non differiscono di molto. Problematici sono invece gli individui sopra la bisettrice, in quanto per questi il peso finale risulta maggiore del peso iniziale. Colorando gli individui con colori diversi a seconda del gruppo di appartenenza individuato dal 'modello a mistura finita' si osserva come il gruppo verde e il gruppo blu, rispettivamente indicati come gruppo 3 e gruppo 4, sono composti da quegli individui per cui la dieta non ha un gran effetto (blu) o per cui addirittura ha un effetto negativo (gruppo verde). Si vede chiaramente che per gran parte dei soggetti vi è un beneficio della dieta. Come in tutte le situazioni sperimentali e reali ad uno stesso stimolo i soggetti reagiscono in modo diverso, quindi è naturale trovare una variabilità nei risultati che è legata al campione in esame, all'eterogeneità degli individui ma anche alla natura stessa del fenomeno.



-

Gruppi	Proporzione del campione totale	Peso medio iniziale	Perdita di peso medio (KG)	Rapporto di Mascolinità	Età media	Variazione Percentuale Peso (Rispetto Basale)
Gruppo 1	14%	99.99	19.47	0.26	44.82	19.5%
Gruppo 2	64%	90.08	8.49	0.23	46.00	9.4%
Gruppo 5	6%	104.45	25.67	0.32	42.00	24.6%
Gruppo 4	13%	83.45	1.11	0.24	44.00	1.3%
Gruppo 3	2%	90.29	-3.65	0.30	43.00	-4.0%

Conclusioni

Il campione di dati fornito da Bioimis relativo alla applicazione da parte di clienti del programma alimentare sottoposto ha evidenziato le seguenti caratteristiche: a) la rilevazione e l'imputazione delle risposte dei clienti avviene attraverso il software 'Qlick View'; b) il data base generato è complesso in quanto prevede la rilevazione di numerose variabili di diversa natura: anagrafiche, esami di laboratorio comunicate dal cliente; c) il criterio di rilevazione dei dati genera diversi tipi di errori nelle risposte che possono essere ad es. di misurazione mancate risposte; d) è necessario introdurre un sistema che controlli la qualità dei dati inseriti nel data base soprattutto che siano imputate le date corrispondenti ad ogni misurazione del peso.
Come primo obiettivo l'analisi si è rivolta alla individuazione delle variabili che con maggiore probabilità generano mancate risposte al peso.
L'applicazione delle due metodologie: Segmentazione binaria e Modello logistico ha evidenziato una coerenza nella individuazione delle variabili che interagiscono con i 'missing'. Infatti, escludendo la prima variabile discriminate -durata di permanenza nella 'forma ideale'"- che è correlata positivamente con le mancate risposte, risulta che è il 'tipo di contratto Live' ad essere più influente sui 'missing' connesso ad una età di ingresso in FI minore di 50 anni.
Il secondo obiettivo è stato rivolto alla valutazione dell'efficacia della dieta proposta dall'accademia alimentare. Lo studio osservazionale condotto su un campione senza mancate risposte alla variabile 'peso' è stato condotto in un primo momento attraverso la 'cluster analysis'. Con questa metodologia si è cercato di individuare gruppi di clienti che abbiano conseguito risultati diversi, delle variabili quantitative presenti nella base dati e principalmente connesse alla diminuzione del peso.
Utilizzando il criterio di Ward che massimizza la diversità tra i gruppi individuati e ne ottimizza l'omogeneità interna, si sono individuati quattro gruppi di clienti diversi all'ingresso della forma ideale: appena in sovrappeso, in sovrappeso, in accentuato sovrappeso, in forte sovrappeso. Ciascun caratterizzato da diversi valori delle età medie e durata nella forma ideale. Relativamente ai valori medi della perdita di peso si è riscontrata la seguente situazione:

- il gruppo degli '**appena in sovrappeso'** presenta una perdita di peso di circa 7kg (circa 9% del peso iniziale);
- il gruppo dei '**sovrappeso**' presenta una perdita di peso di circa 11 Kg (circa 12% del peso iniziale);
- il gruppo dei' **sovrappeso accentuato**' presenta una perdita di peso di circa 11 Kg (circa 12% del peso iniziale
- il gruppo dei '**forti sovrappesi'** presenta una perdita di peso di circa 15 kg (circa 12% del peso iniziale);

Complessivamente utilizzando i valori medi i 2684 clienti hanno perso 28652 kg che corrispondono in media a 10,67 kg nel periodo medio di FI di 74 giorni.
Una seconda analisi osservazionale è stata condotta tenendo conto dell'andamento del peso ai diversi momenti di osservazione, il genere la durata del tempo di osservazione l'età di ingresso in forma ideale (inizio della dieta) e stato di salute (sano è la modalità di riferimento). I risultati evidenziano 5 gruppi di cui: tre di essi presentano una diminuzione di peso che varia da 25 a 8 kg. con una variazione da 24% a 9% rispetto al peso iniziale. Un gruppo presenta una lieve diminuzione

del peso sia in termini assoluti (1kg). che percentuali (circa 1%). Con questa analisi si evidenzia la presenza di un gruppo, circa il 2% del campione, che non ha ottenuto benefici dal regime alimentare: un aumento di circa il 4% del loro peso che corrisponde a circa 3 kg.
In conclusione le due analisi adottate per valutare l'efficacia della dieta alimentare, proposta da Bioimis, evidenziano che in gran parte dei casi si registra una diminuzione media del peso con diverse tarature. Ovviamente, come ogni fenomeno statistico, esistono dei casi che registrano un comportamento diverso dalla media in senso negativo. In questo campione sono solo circa il 2%.