# Big Data in Health Care

Federico Moiraghi - 799735, Pranav Kasela - 846965

A.A. 2019/2020

**Sommario**

Obiettivo del presente Progetto è di fornire un modello predittivo che riesca a riconoscere i tumori omogenei da quelli eterogenei. Infatti l'eterogeneità del tumore rappresenta una difficoltà aggiuntiva nella fase di trattamento, rendendolo maggiormente resistente alle cure. Il modello risultante dal presente Progetto sarà facilmente interpretabile da un esperto di dominio, in modo tale da supportare le sue decisioni e non sostituirsi completamente ad esso.

## Indice

# 1 Introduzione

Si è deciso di utilizzare per il presente Progetto il linguaggio di programmazione `R`, dato che il suo uso all'interno di ospedali e centri di ricerca è in forte crescita. Grazie alla libreria `RNifti` è possibile importare i file `.nii` (contenenti le lesioni oncologiche), facilitando sia l'analisi esplorativa sia la successiva costruzione del modello.

Caricate le immagini (si ha un esempio con figura 1), si nota che queste rappresentano la lesione già segmentata. Prima dell'analisi vera e propria è necessario ritagliare l'immagine, eliminando il valore dei *voxel* esterni alla lesione: questi infatti hanno valore nullo, mentre, per aumentare l'efficienza del *workflow*, si preferisce eliminare tali valori rendendoli non definiti (`NA`).

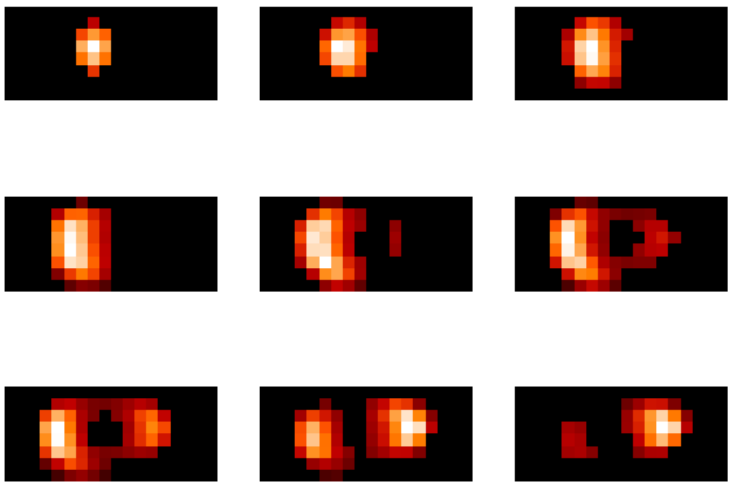

Figura 1: Esempio di immagine. Essendo una figura tridimensionale, si rappresenta la profondità con più immagini.

Mentre le dimensioni dei *voxel* sono fisse per tutti i files (tabella 1), l'immagine è ritagliata sulla lesione in modo specifico, tralasciando le parti adiacenti: le dimensioni variano in base alla dimensione della lesione stessa.

| | x | y | z |
|---|---|---|---|
| 0 | 2.734 | 2.734 | 2.734 |

Dunque sarà importante estrarre delle *features* che non dipendano dalla dimensione dell'immagine ma tengano conto di possibili variazioni. Questo approccio comporta una serie di vantaggi, primo tra tutti la modularità del *workflow*: è possibile così prevedere la variabile risposta avendo a disposizione sia un'immagine già segmentata sia effettuando la segmentazione *on-the-fly* tramite semplici algoritmi a soglia.

## 2 Estrazione delle *features*

L'estrazione delle *features* mappa le immagini in uno spazio di dimensionalità molto minore, rendendo più semplice l'analisi dato il numero esiguo di dati a disposizione. Infatti, un qualsiasi algoritmo di *machine learning* ha bisogno di un numero significativo di dati per "apprendere" in modo *data-driven* cosa utilizzare nell'analisi.

Per selezionare le *features* da utilizzare, si è preso spunto da [1][1]: si usano i momenti dal primo fino al quarto (media, varianza, asimmetria e curtosi) dei valori dei *voxel* che raffigurano la lesione oncologica. A questi però si è deciso di aggiungere anche il volume della lesione (`volume`) e della sfera equivalente (`sphere`), così come anche la superficie (`area`) della lesione (figura 1). Purtroppo, non essendo stata specificata l'unità di misura dei *voxel*, non è possibile stabilirla nemmeno per le *features* ricavate; tuttavia si presume che il proprietario dei dati sia in possesso di tale informazione e che quindi riesca ad attribuire maggiore semantica.

| VoxelNum | Maximum3DDiameter | MajorAxisLength | Sphericity | MinorAxisLength | Surfac |
|---:|---:|---:|---:|---:|---|
| 0.018 | -0.695 | -0.673 | 1.022 | -0.266 | |
| -0.661 | 0.436 | 0.812 | -0.185 | -1.274 | |
| -0.326 | -0.719 | -0.751 | 0.736 | -0.525 | |
| 0.485 | 0.356 | 0.028 | 0.787 | 0.23 | |
| -0.797 | -0.986 | -0.977 | 0.107 | -0.913 | |
| 0.493 | -0.236 | -0.338 | 1.258 | 0.063 | |

Tabella 1: Esempio di *features* estratte per le singole immagini.

La matrice dei dati risultante è quindi di dimensioni $44 \times 24$ (si esclude la variabile risposta), indipendentemente dalla dimensione delle immagini di partenza o dalla loro risoluzione.

[1]Gli autori usano i primi quattro momenti per stimare la differenza di eterogeneità di tumori alla cervicale nel tempo, a seguito di un trattamento.

Prima di procedere con la costruzione del modello, si preferisce effettuare una rapida analisi esplorativa sulla nuova matrice per verificare quali varibili includere.

Il primo metodo della feature selection e' usando il test di Mann-Whitney, equivalente non parametrico del t-test.

| Maximum3DDiameter | MajorAxisLength | Sphericity | MinorAxisLength | SurfaceArea | Kur |
|---|---|---|---|---|---|
| 0.011 | 0.012 | 0.003 | 0.003 | 0.016 | 0 |

Tabella 2: p-values delle varibili accetate dal test di Mann-Whitney.

Una volta scelto le variabili importanti vediamo la loro correlazione, per escluderle, questo e' necessario per evitare problemi di multicollinearita'.

Dalla figura 2, deduciamo subito che il Maximum e Variance sono correlati a piu' di una variabile, quindi e' meglio escluderli.

# 3 Costruzione del modello

Essendo la variabile risposta binaria (tumore *omogeneo* o *eterogeneo*, rispettivamente 0 o 1), e volendo costruire un modello facilmente interpretabile per un esperto di dominio, si effettua una semplice regressione logistica.

| | average | $IDC_{99}$ |
|---|---|---|
| accuracy | 0.863 | 0.045 |
| precision | 0.92 | 0.052 |
| recall | 0.853 | 0.07 |
| $f_1$ | 0.873 | 0.046 |

La selezione delle *features* è effettuata tramite procedimento *stepwise* (partendo dal modello pieno ed eliminando le variabili superflue, ma con la possibilità, a ogni iterazione, di reinserirle). Si è deciso di rimuovere alcune variabili a prescindere:

- `sd` (la varianza della distribuzione della lesione), in quanto fortemente correlata con `mean` (0.93) e di più difficile interpretazione;
- `area` (la superficie della lesione), in quanto la sua stima è approssimativa e risulta essere eccessivamente correlata ad altri regressori, inquinando eccessivamente la qualità dei dati;
- `sphere` (il volume della sfera equivalente), data la sua forte correlazione con la variabile `volume` (0.86) e la più difficile interpretabilità.

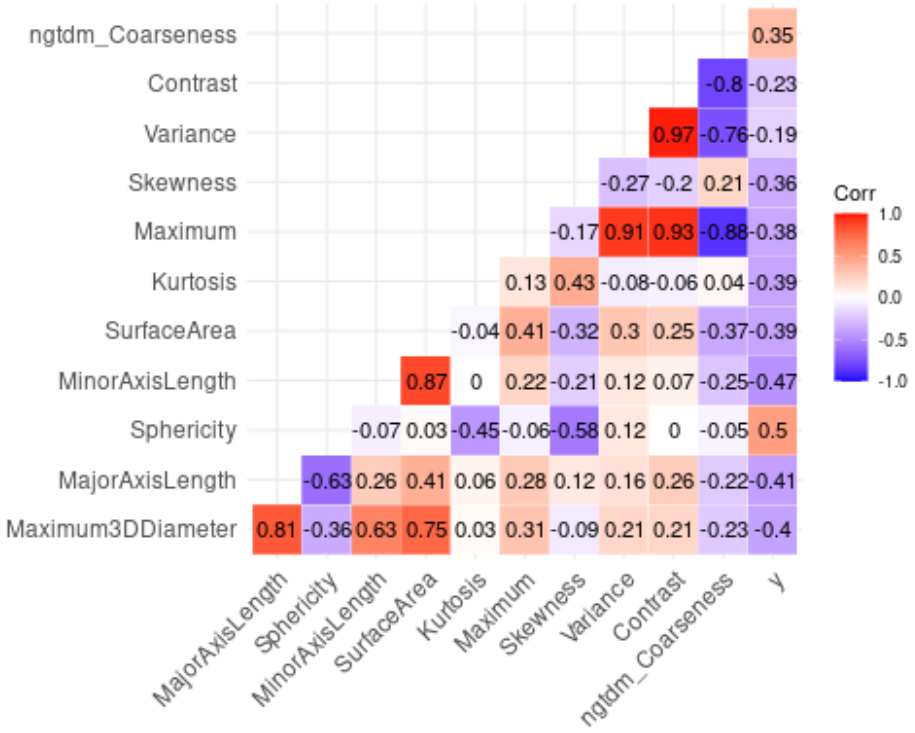


Figura 2: Correlogramma delle *features* estratte.

Avendo a disposizione pochi dati, il procedimento è effettuato con l'indice AIC, che considera la capacità di generalizzazione del modello complessivo risultante (salvo poi verificare le *performance* su un *test set* composto da dati nuovi, rappresentante circa il 20% di quelli totali). Alla fine del procedimento, il modello risultante comprende solo tre regressori (più l'intercetta), come mostrato in tabella 3.

| | Stima | p-value |
|---|---|---|
| (Intercept) | -3.721629 | 0.029311 |
| SurfaceArea | -10.495655 | 0.007634 |
| Kurtosis | -8.183723 | 0.019343 |
| Skewness | -7.248026 | 0.015959 |

Tabella 3: Stima dei coefficienti del modello e loro significatività.

Si noti come i coefficienti maggiormente significativi siano l'asimmetria `sk` e il volume `volume`: la probabilità che il tumore sia eterogeneo è tanto maggiore quanto più grande è la lesione e quanto più pesante è la coda positiva della distribuzione. Si può ipotizzare infatti che questa coda positiva sia costituita da sotto-componenti particolarmente aggressivi del tumore, quindi "ghiotti" di traccianti e di conseguenza maggiormente visibili nell'immagine.

| | heterogeneous | homogeneous |
|---|---|---|
| heterogeneous | 5 | 0 |
| homogeneous | 0 | 4 |

Tabella 4: Matrice di confusione del modello di regressione logistica per il *test set*; sulle righe le previsioni e sulle colonne i valori reali.

Come si può notare nella matrice di confusione (figura 4), il modello ha commesso un solo errore catalogando come eterogenea una lesione omogenea.

| | |
|---|---|
| accuracy | 1 |
| precision | 1 |
| recall | 1 |
| $f_1$ | 1 |

# 4 Conclusioni

Con questo Progetto si è costruito un modello statistico efficace e facilmente interpretabile da un esperto di dominio per prevedere l'eterogeneità del tumore. Si è visto che, estrapolando dall'immagine dei semplici valori indice, è

possibile costruire un modello indipendente dalla dimensione dell'immagine o dalla sua risoluzione.
A livello matematico si potrebbe aumentare la prestazione del modello stimando i parametri con un numero maggiore di dati; tuttavia, in ambito medico, questo non è sempre possibile (anche perché, come espresso in [1], è possibile verificare l'eterogeneità del tumore solo in modo invasivo o con autopsia). Inoltre si potrebbero utilizzare nuove *features*, soprattutto se utili ai fini della ricerca medica.

# Riferimenti bibliografici

[1] Stephen Bowen, William Yuh, Daniel Hippe, Wei Wu, Savannah Partridge, Saba Elias, Guang Jia, Zhibin Huang, George Sandison, Dennis Nelson, Michael Knopp, Simon Lo, Paul Kinahan, and Nina Mayr. Tumor radiomic heterogeneity: Multiparametric functional imaging to characterize variability and predict response following cervical cancer radiation therapy. *Journal of Magnetic Resonance Imaging*, 47, 10 2017.