Big Data in Health Care

Federico Moiraghi - 799735 & Pranav Kasela - 846965

A.A. 2019/2020

Sommario

Obiettivo del presente Progetto è fornire due modelli predittivi che riconoscano 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 primo modello, *supervised*, sarà facilmente interpretabile da un esperto di dominio, in modo tale da supportare le sue decisioni senza sostituirsi completamente ad esso. Tale modello sarà poi confrontato con uno *unsupervised*, che sottolinea le analogie tra i singoli casi.

Indice

1 Introduzione

Si è deciso di sviluppare il presente Progetto coi linguaggi di programmazione Python ed R, data la forte crescita del loro uso sia in ambiente accademico che produttivo. Il primo è usato soprattutto per l'estrazione delle *features* dalle immagini, grazie alla libreria `pyradiomics` che offre numerosi algoritmi; mentre il secondo per l'analisi dei dati, data l'ampia scelta di modelli di *machine learning*.

Caricate le immagini (si ha un esempio con figura 1), si nota che queste rappresentano la lesione già segmentata. Non si ritiene dunque necessaria alcuna forma particolare di *pre-processing* sull'immagine.

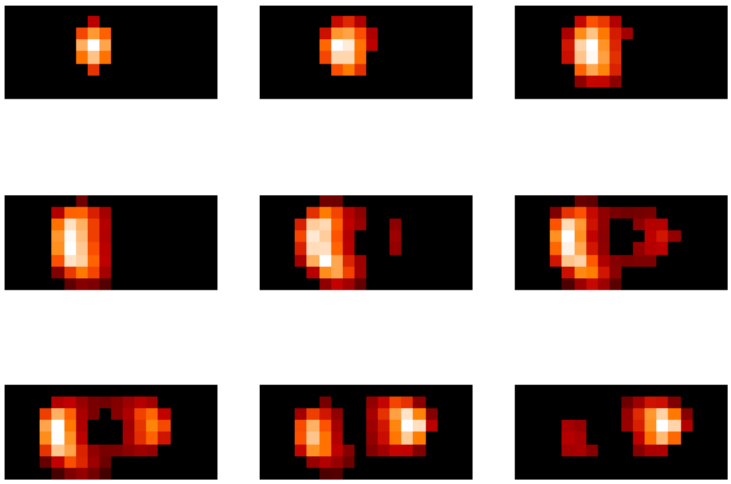

Figura 1: Esempio di immagine. Essendo una figura tridimensionale, si rappresenta la profondità con più immagini.

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

	voxel size
x	2.734
y	2.734
z	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, riducendo potenzialmente il tempo della diagnosi.

2 Estrazione delle *features*

L'estrazione delle *features* mappa le immagini in uno spazio di dimensionalità molto minore e di dimensioni fisse, 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] e [2]: si usano *features* estratte direttamente dall'immagine, quali superficie della lesione o la sua sfericità, accompagnate da indici statistici più semplici quali i momenti di ordine dal primo fino al quarto (media, varianza, asimmetria e curtosi) dei valori dei singoli *voxel* (tabella 1).

VoxelNum	Maximum3DDiameter	MajorAxisLength	Sphericity
0.485	0.356	0.028	0.787
-0.295	-0.719	-0.747	0.85
0.458	0.118	-0.106	0.956
-0.341	-0.6	-0.021	-1.781
-0.326	-0.719	-0.751	0.736
-0.739	-0.14	0.062	-1.256

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

Tuttavia il numero di *features* estratte è ancora elevato rispetto al numero di dati a disposizione (la matrice di *input* ha dimensioni 44×20). Si effettua dunque una prima selezione delle *features* tramite il test di Mann-Whitney, equivalente non parametrico del t-test (i risultati sono riassunti nella tabella 2), per escludere le variabili che, prese singolarmente, hanno una significatività minore del 95%.

[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.

Kurtosis	0.002	**
Sphericity	0.003	**
MinorAxisLength	0.003	**
Maximum	0.008	**
Maximum3DDiameter	0.011	*
MajorAxisLength	0.012	*
SurfaceArea	0.016	*
Variance	0.025	*
Coarseness	0.034	*
Skewness	0.04	*
MeanAbsoluteDeviation	0.051	.
Mean	0.079	.
Entropy	0.102	
Median	0.137	
Contrast	0.157	
Minimum	0.204	
VoxelNum	0.219	
VoxelVolume	0.219	
Flatness	0.291	
SurfaceVolumeRatio	0.878	

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

Effettuata questa prima cernita, si riduce ulteriormente il numero di *features*, in modo tale da evitare multi-collinearità tra le variabili, rispettando così le premesse del modello lineare.

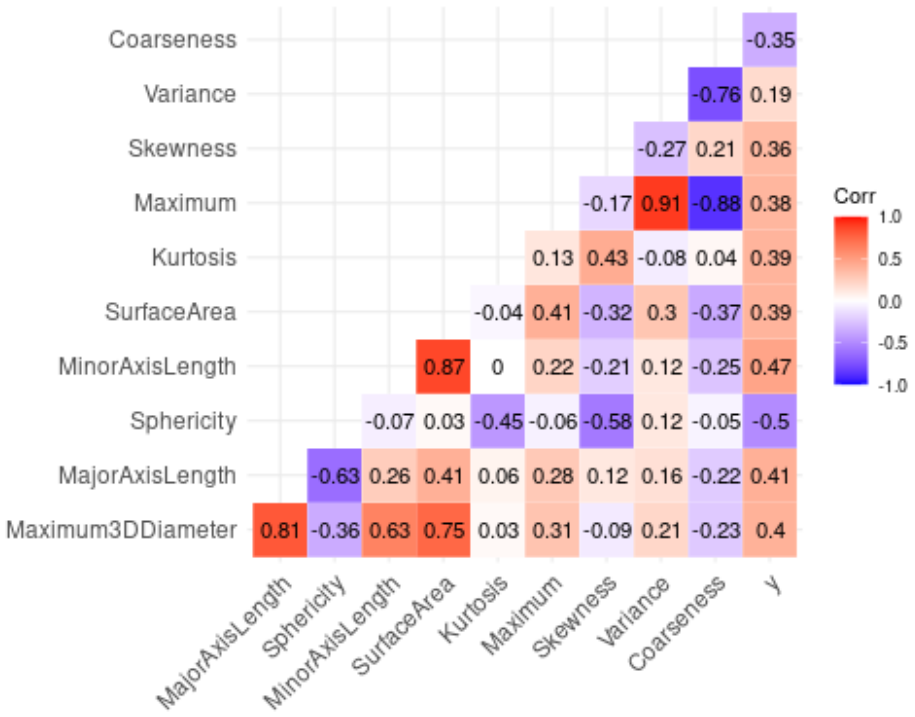

Figura 2: Correlogramma delle *features* estratte.

Dal correlogramma (figura 2) si deduce quali variabili escludere (`Maximum`, `Variance`, `Maximum3DDiameter`, `MinorAxisLength`, `Contrast` e `Sphericity`): la matrice risultante ha una dimensionalità ridotta (44×5), adeguata per la costruzione del modello.

Nella figura 3, viene mostrata la distribuzione di densità delle variabili accettate, condizionata alla tipologia di lesione, in modo da vedere graficamente la differenza nella distribuzione.

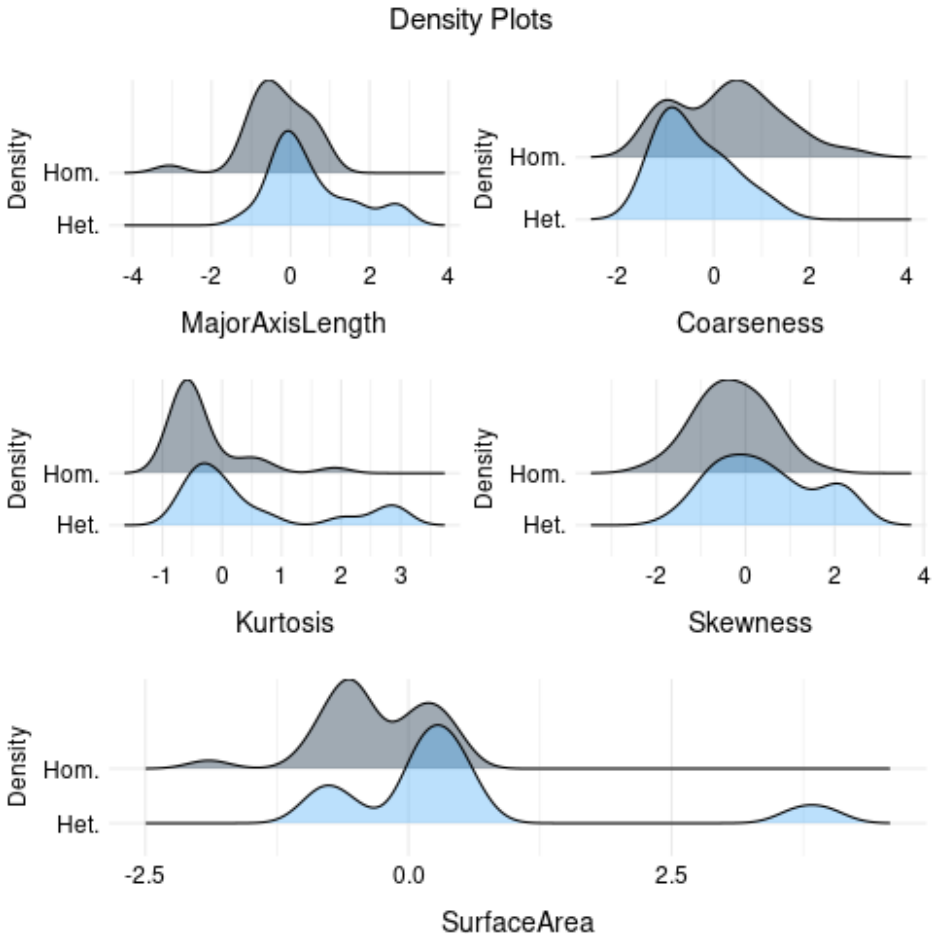

Figura 3: *Density plot* delle variabili scelte per tipo di lesione.

3 Modello *supervised*

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.

La selezione delle *features* è effettuata tramite procedimento *stepwise* usando l'indice BIC[2], con possibilità di re-immissione. Il numero di variabili significative si riduce quindi a tre: `SurfaceArea`, `Kurtosis` e `Skewness` (riassunti nella tabella 3 coi rispettivi p-value).

	Stima	p-value
(Intercept)	4.295873	0.014587
SurfaceArea	11.899879	0.005449
Kurtosis	9.842963	0.008876
Skewness	8.655905	0.007367

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

Il modello, come previsto, sottolinea la correlazione tra la superficie della lesione e la sua eterogeneità: un tumore eterogeneo, infatti, ha spesso una forma irregolare e dunque una superficie maggiore. Inoltre si nota come anche curtosi e asimmetria positiva siano relazionate con la probabilità di eterogeneità: se un tumore è composto da componenti più "ghiotte" (e dunque aggressive), i rispettivi *voxel* risultano maggiormente visibili e quindi entrambi gli indici aumentano.

Le prestazioni del modello sono calcolate col sistema *iterated holdout*, effettuando 30 iterazioni casuali dividendo i dati 80% *train set* e 20% *test set*, così da avere stime robuste dei parametri e un intervallo di confidenza sufficientemente ristretto. La media degli indici di bontà è riportata nella tabella 4 assieme al rispettivo intervallo di confidenza al 99%.

	average	IDC_{99}
accuracy	0.867	0.059
precision	0.807	0.117
recall	0.835	0.118
f_1	0.813	0.093

Tabella 4: Performance del modello supervisionato con intervallo di confidenza al 99%.

[2]L'indice BIC rispetto all'indice AIC penalizza maggiormente l'inserimento di una nuova variabile con un numero ridotto di osservazioni.

La tabella 5 mostra la *confusion matrix* (del *test set*) dell'ultima iterazione: il modello ha una buona capacità predittiva.

	heterogeneous	homogeneous
prediction: heterogeneous	5	0
prediction: homogeneous	0	4

Tabella 5: Matrice di confusione del modello di regressione logistica per il *test set*; sulle righe le previsioni e sulle colonne i valori reali.

4 Modello *unsupervised*

Dopo aver standardizzato le variabili, si eliminano quelle con coefficiente di correlazione molto alto, per non introdurre troppo rumore nel modello. Si ipotizza così che i tumori tra loro simili appartengano alla stessa classe.

Le immagini di tumori quindi sono collocate in uno spazio vettoriale in base al risultato della *Principal Component Analysis* (PCA): si selezionano così le prime 5 componenti, che spiegano almeno il 5% della varianza della distribuzione. Così, operando su una matrice di dimensioni ridotte, si riduce la quantità di rumore data dall'elevato numero di variabili (a cui si esclude la variabile risposta `y`, usata poi per calcolare la bontà del modello). Dalla figura 4 infatti si evince che all'aumentare del numero di componenti considerate la percentuale di varianza spiegata dalla componente decresce: la soglia del 5% è un compromesso tra il segnale colto dal modello e la sua complessità (per i dettagli vedere la tabella 6). Le 5 componenti scelte spiegano complessivamente circa il 90% della varianza totale.

	PC1	PC2	PC3	PC4	PC5
Standard deviation	1.912	1.6	1.391	1	0.839
Proportion of Variance	0.332	0.233	0.176	0.091	0.064
Cumulative Proportion	0.332	0.565	0.741	0.832	0.896

Tabella 6: Alcune statistiche sulle prime componenti principali.

Nello spazio della PCA si effettua un raggruppamento usando l'algoritmo DBScan, basato sulla densità delle osservazioni. La figura 5 suggerisce un parametro $\varepsilon \in (2.5, 3)$ (lo si vede dal salto nella figura con 5-NN), si opta per il valore medio: questa configurazione sarà usata per la costruzione del modello.

Nonostante le buone *performance* del modello (riassunte nella tabella 7), si nota che il secondo *cluster* C_1 contiene un numero non indifferente, pur

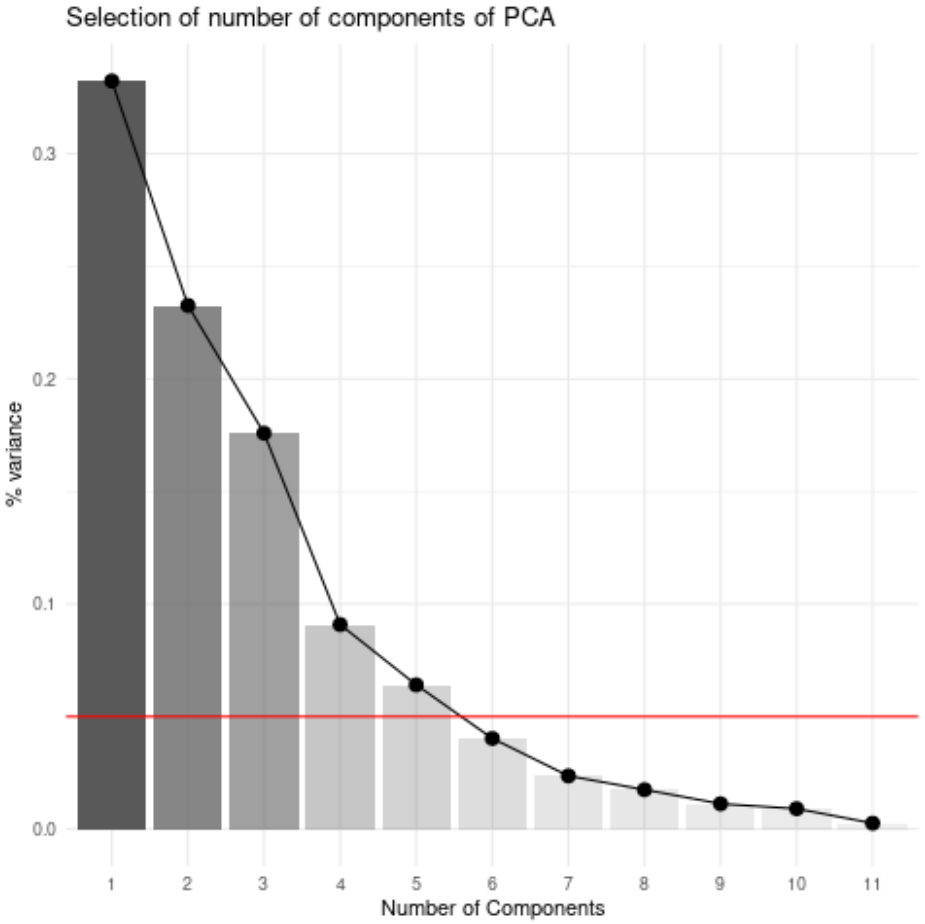

Figura 4: Andamento della varianza spiegata dal modello all'aumentare del numero di componenti della PCA.

accuracy	0.773
precision	0.5
recall	0.9
f_1	0.643

Tabella 7: Indici di bontà per la clusterizzazione con DBScan.

	heterogeneous	homogeneous
C_0	9	1
C_1	9	25

Tabella 8: Distribuzione delle immagini all'interno dei *clusters*.

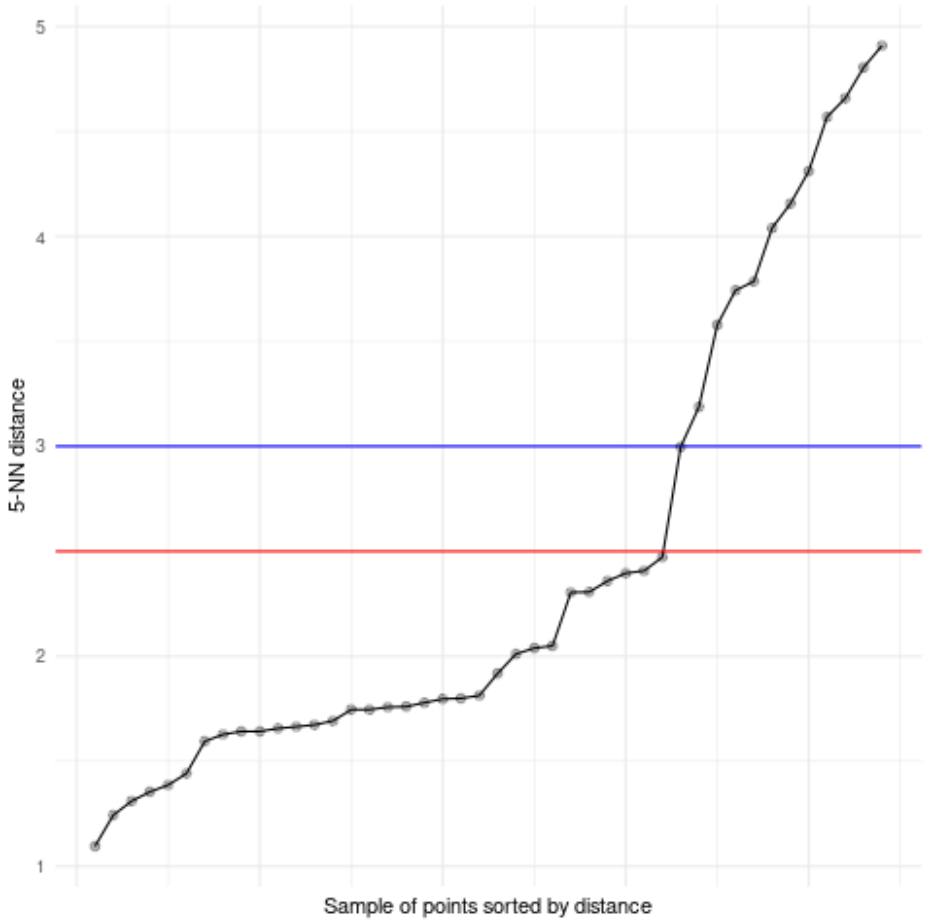

Figura 5: Scelta del valore ε per DBScan.

rimanendo ridotto, di immagini eterogenee (tabella 8). L'algoritmo è riuscito a individuare un *cluster* (C_0, il cluster di elementi rigettati) ben definito, considerando la variabile risposta.

Si tenta un altro approccio, con l'algoritmo *HK-means*, versione gerarchica del ben più noto *K-means*. L'algoritmo è quindi testato con un numero di *cluster* k da 2 a 15, calcolando per ciascuno la distanza nei gruppi (*distance between*). La figura 6 mostra graficamente il procedimento: si sceglie $k = 5$ per evitare *overfitting* dei dati, siccome il tasso di miglioramento per $k > 5$ decresce fortemente. La bontà del raggruppamento (intesa come capacità predittiva) è invece riassunta nella tabella 9.

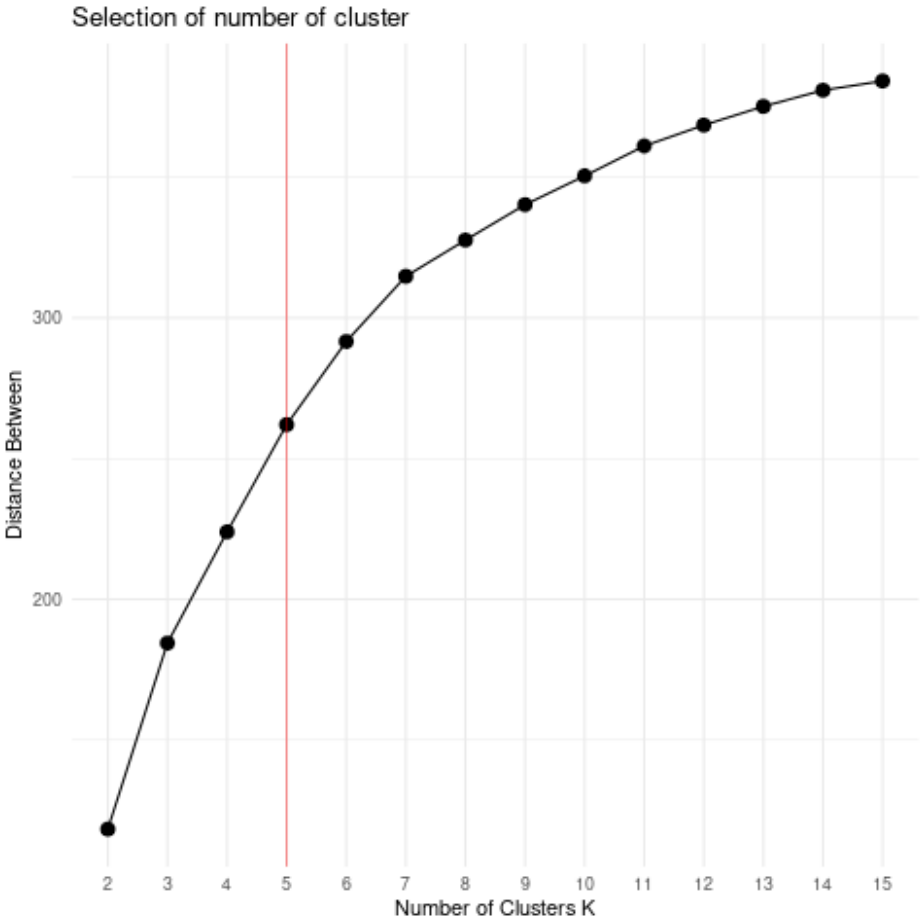

Figura 6: Variazione della distanza *between* all'aumentare del parametro k.

Il modello è riuscito ad individuare 3 cluster ben distinti, 2 per le prove eterogenee (C_4 e C_5) e 1 per le prove omogenee (C_1). Il *cluster 2* e *cluster 3* invece hanno un contenuto ambiguo, che comprende osservazioni di entrambi

accuracy	0.786
precision	0.438
recall	1
f_1	0.609

Tabella 9: Indici di bontà per HK-Means con $k = 5$.

	heterogeneous	homogeneous
C_1	0	10
C_2	5	8
C_3	4	8
C_4	7	0
C_5	2	0

Tabella 10: Il modello ha identificato tre *cluster* definiti (C_1, C_4 e C_5), considerando la variabile risposta;

i gruppi. Questo si verifica anche a causa della similarità tra distribuzione delle variabili condizionati alla lesione (figura 3).

Dopo aver provato sia HK-Means che DBScan, si deduce che la capacità predittiva dei modelli *unsupervised*, paragonati a quello *supervised* presentato precedentemente, è nettamente inferiore. Inoltre, lavorando nello spazio delle componenti principali, l'interpretabilità del modello risulta difficile anche per un esperto di dominio.

5 Conclusioni

Con questo Progetto si è costruito un modello statistico *supervised* efficace e facilmente interpretabile da un esperto di dominio per prevedere l'eterogeneità del tumore. Per costruirlo è stato sufficiente estrapolare dalle immagini segmentate delle semplici *features*, veloci da calcolare e facili da interpretare. Si è quindi confrontato questo modello con uno *unsupervised*, confermando la superiorità del primo sia per bontà di previsione sia per facilità di interpretazione.

Per migliorare il modello si potrebbe, a livello teorico, usare un maggior numero di dati per stimare i parametri e per selezionare le *features* da includere; tuttavia questo non è sempre possibile in ambito medico, data la forte difficoltà e l'alto costo nell'ottenere una più grande quantità di dati. Inoltre, con un numero maggiore di dati è possibile utilizzare modelli più complessi che considerino anche interazioni tra le variabili o *pattern* non lineari, senza rischiare di perdere capacità di generalizzazione.

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.

[2] Francesca Gallivanone, Matteo Interlenghi, Daniela D'Ambrosio, Giuseppe Trifirò, and Isabella Castiglioni. Parameters influencing pet imaging features: a phantom study with irregular and heterogeneous synthetic lesions. *Contrast Media & Molecular Imaging*, 2018:1–12, 2018.