# Data Mining

## Fat prediction on steak images

Francesco Franco, 162167

Daniele Tortoli, 153371

# Indice

# Presentazione del dataset

Il dataset è composto da due immagini (jpg, RGB) di bistecche.

**A)** Steak 1 (500 x 348px)

**B)** Steak 2 (431x341px)

# Obiettivi dell'analisi (1)

- Analisi Esplorativa: utile per studiare il dataset andando a scomporre l'immagine nei suoi **tre** canali (R,G,B) osservandone la distribuzione dell'intensità dei pixel.

- PCA esplorativa: preprocessing, ricerca del miglior tradeoff tra dimensionalità e varianza espressa, scelta di un valore soglia in base agli scores per distinguere la carne dal grasso.

# Obiettivi dell’analisi (2)

- Clustering: test di due algoritmi di clustering e comparazione per osservare quale distingue meglio le nostre classi.

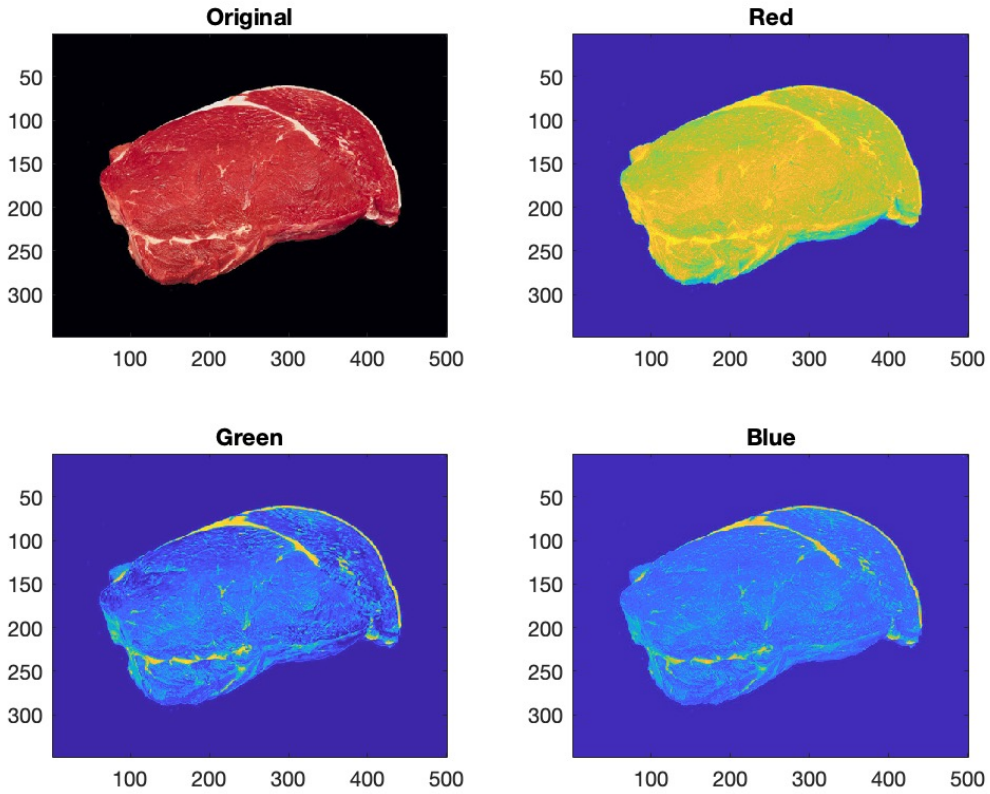
- Applicazione del modello: costruzione modello PCA sulla classe **grasso** e proiezione della seconda immagine su tale modello.

# Analisi esplorativa – Canali RGB

Divisione in canali dell'immagine steak 1.

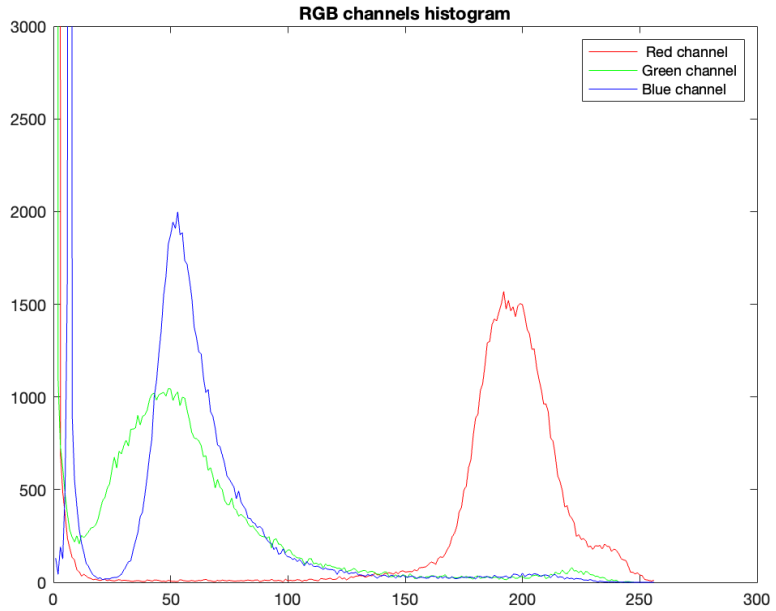
Il canale rosso discrimina meglio la carne dallo sfondo.

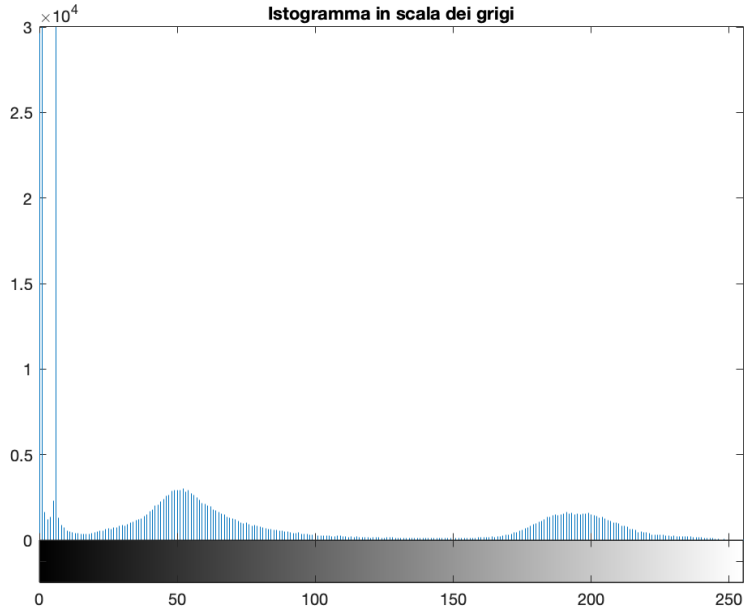
I canali verde e blu distinguono meglio il grasso dal resto.



# Analisi esplorativa - Istogrammi

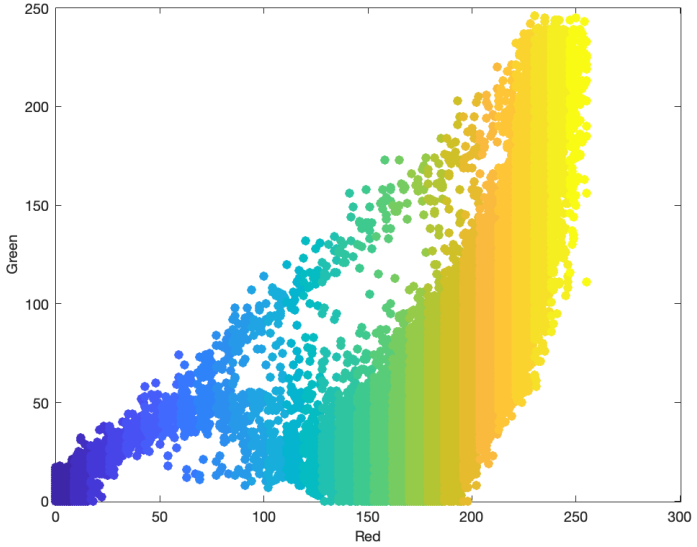
Analizzando gli istogrammi è possibile identificare le regioni che rappresentano lo sfondo, la carne e il grasso.



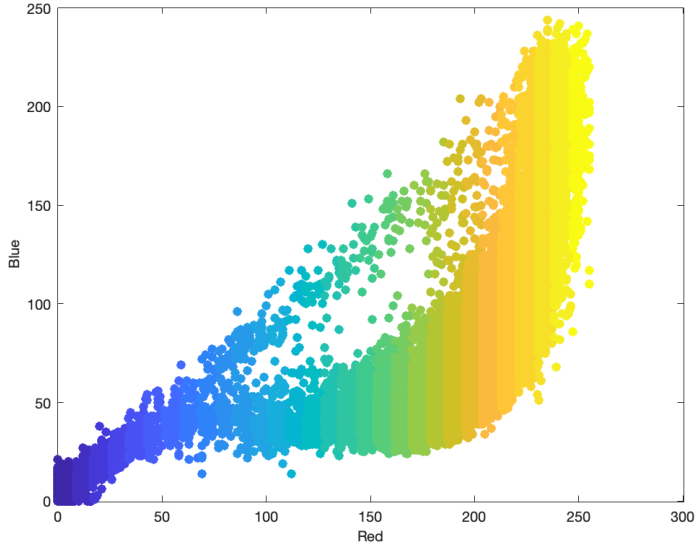



# Analisi esplorativa – Canali a confronto

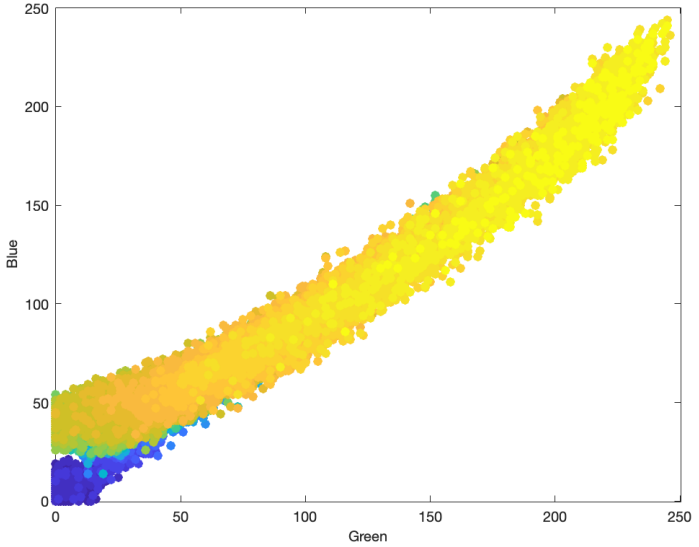
Il canale rosso prevale ed è meno correlato agli altri due.
Vi è invece una maggiore correlazione positiva tra verde e blu.




1) Canali rosso e verde

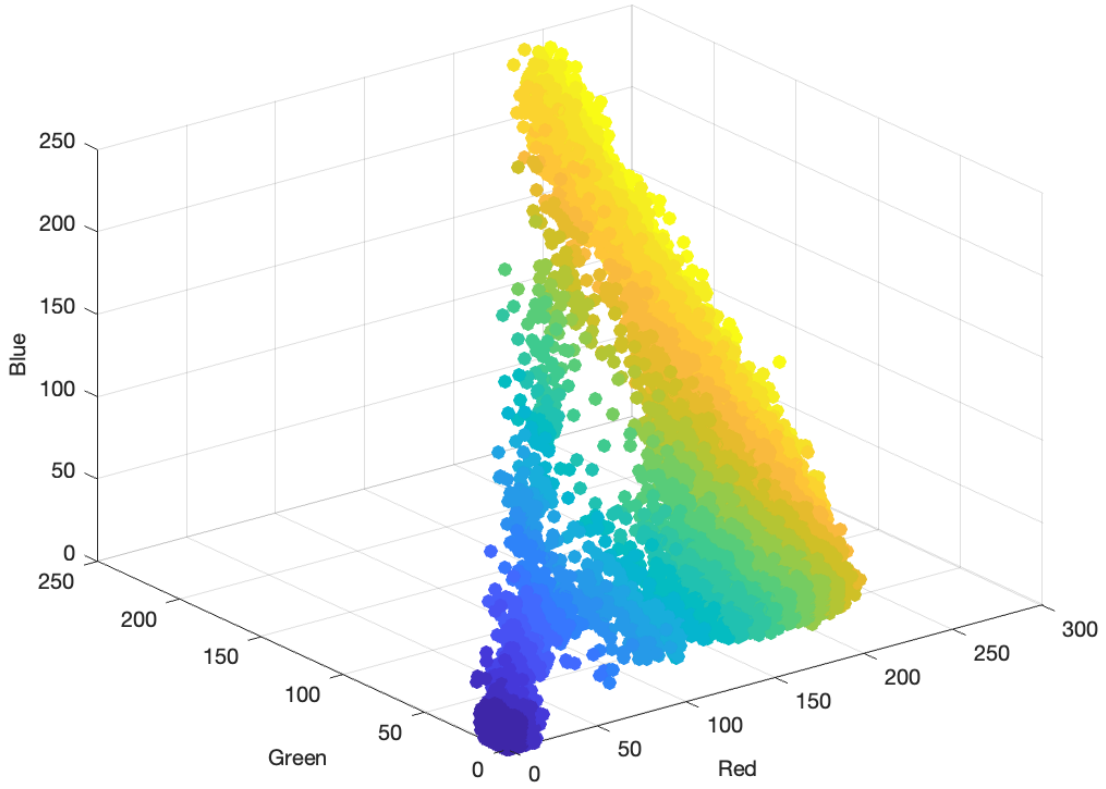



2) Canali rosso e blu




3) Canali verde e blu

# Analisi esplorativa– Scatter 3 canali RGB

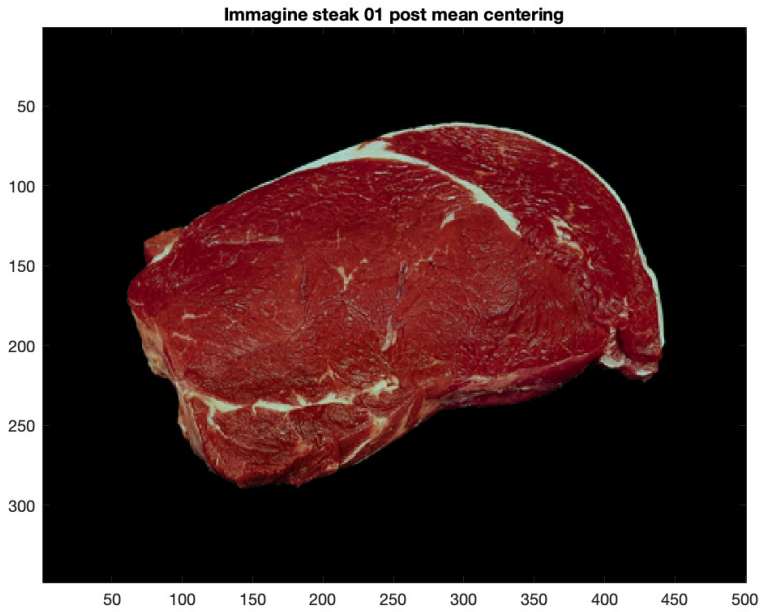


# PCA esplorativa - Preprocessing

Effettuato tramite **mean centering** per canale.

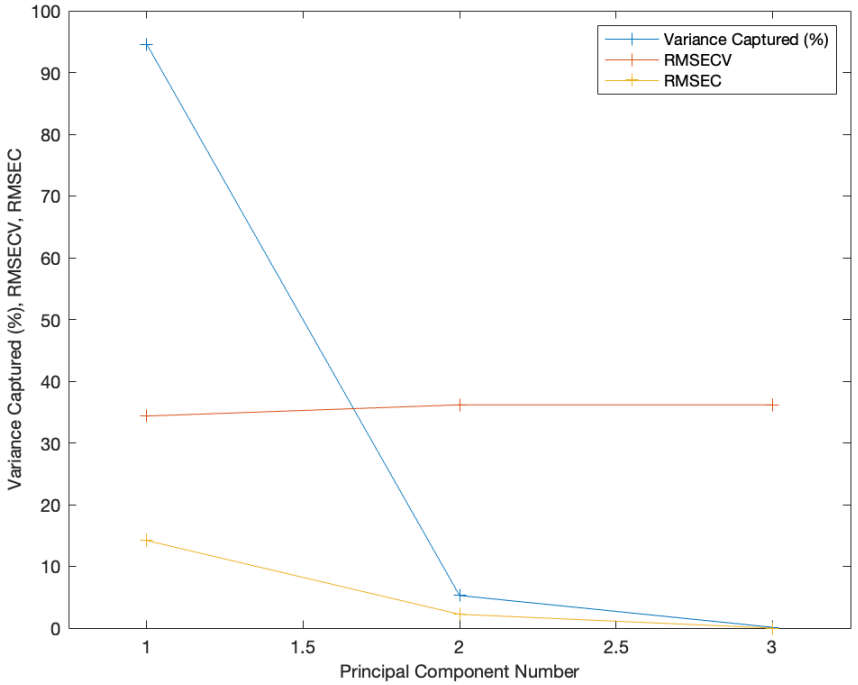
Nello specifico i valori sottratti ai vari canali sono stati i seguenti:

| | R | G | B |
|---|---|---|---|
| **Media** | 67,5762 | 21,0495 | 26,8008 |



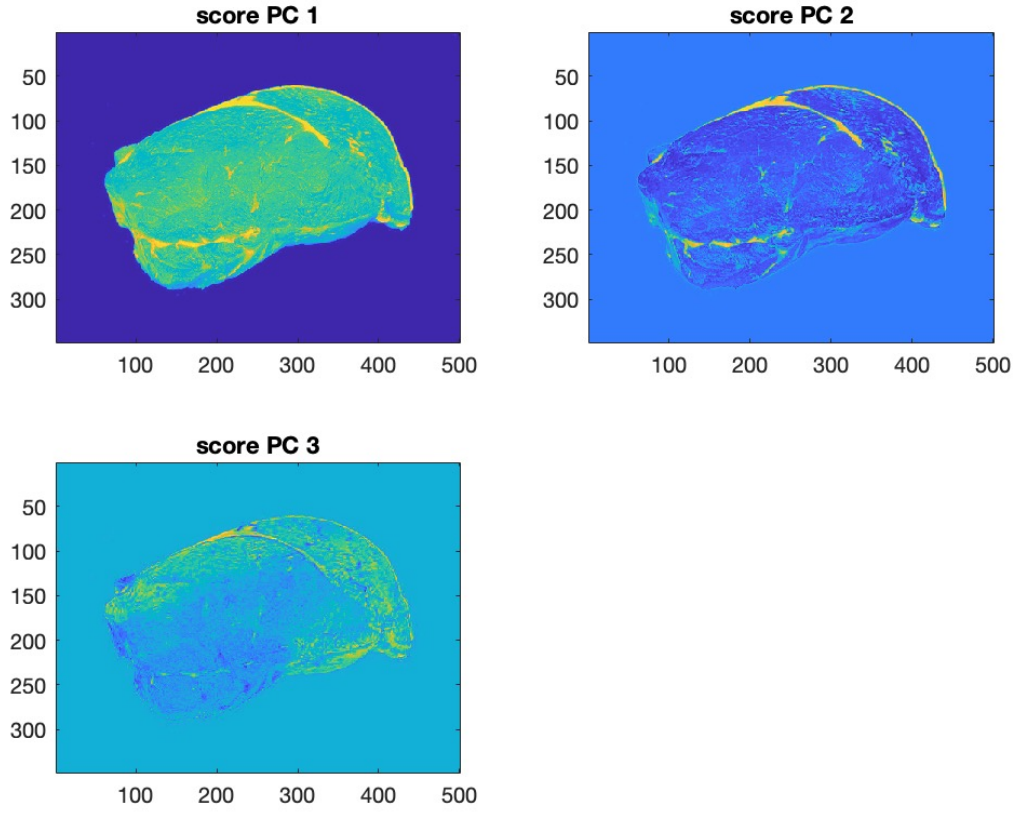
# PCA esplorativa - Scelta delle componenti

Dalla cross validation otteniamo che il numero consigliato di PC è **1**, infatti, essa riesce ad esprimere correttamente quasi il 95% dell'informazione contenuta nei nostri dati.

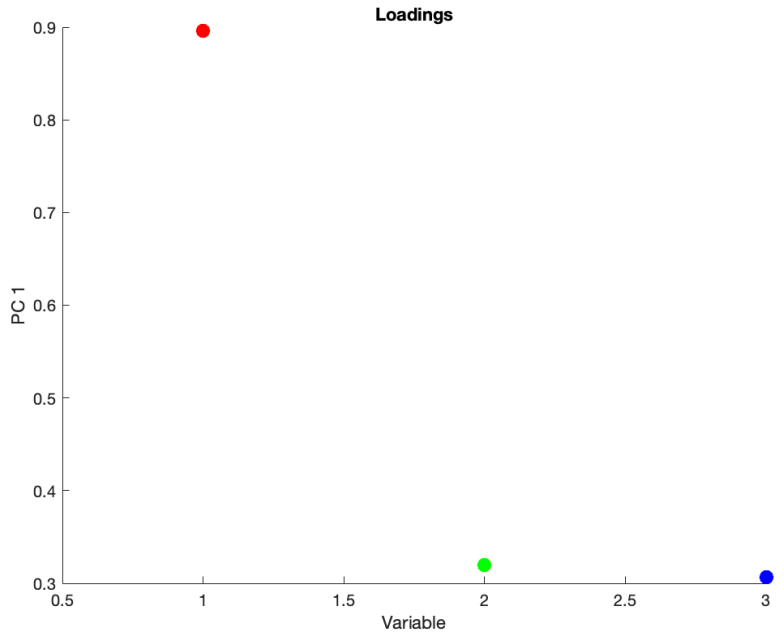


# PCA esplorativa – Score PCA

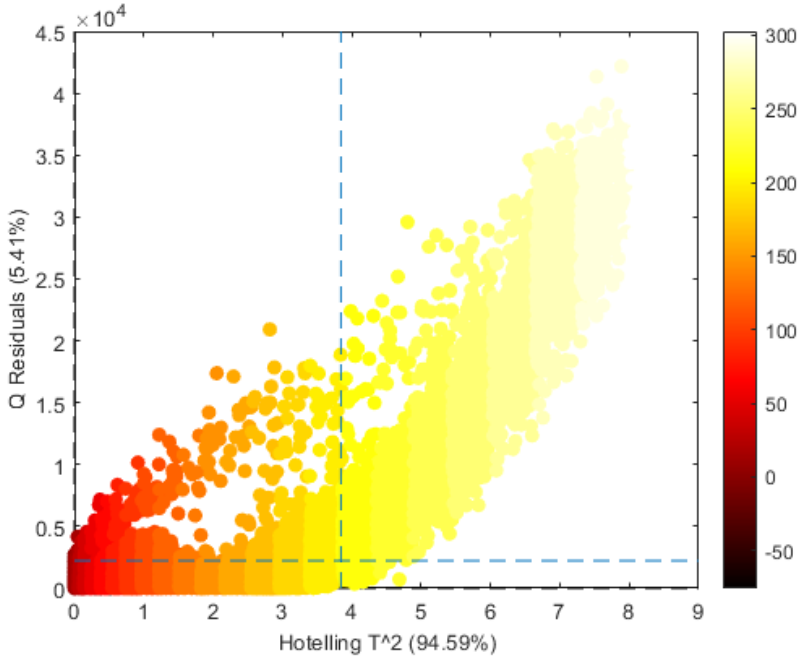
Gli scores della **PC1** sono gli unici in grado di distinguere bene la carne dallo sfondo.



# PCA esplorativa – Loadings e residual




1) I loadings della PC1 evidenziano la netta separazione tra il rosso e gli altri due canali

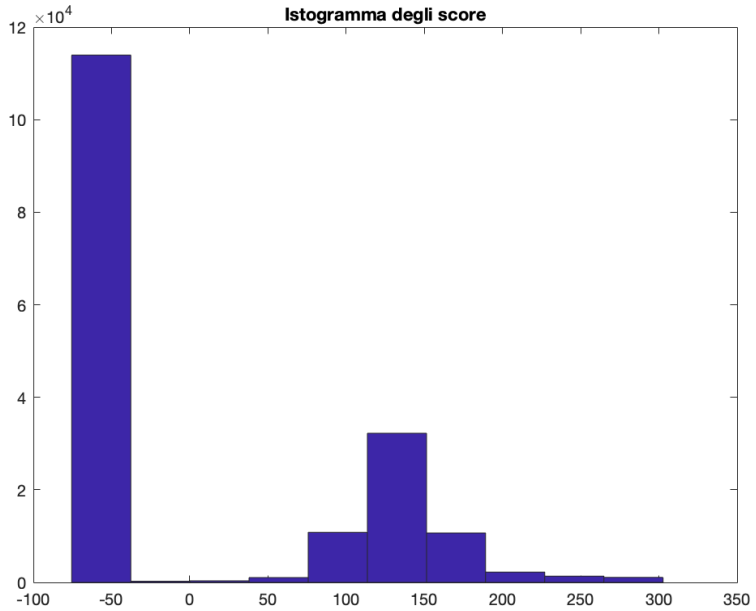



2) Hotelling $T^2$ e Q Residual, colorati in base allo score

# PCA esplorativa – Soglie di valori (1)

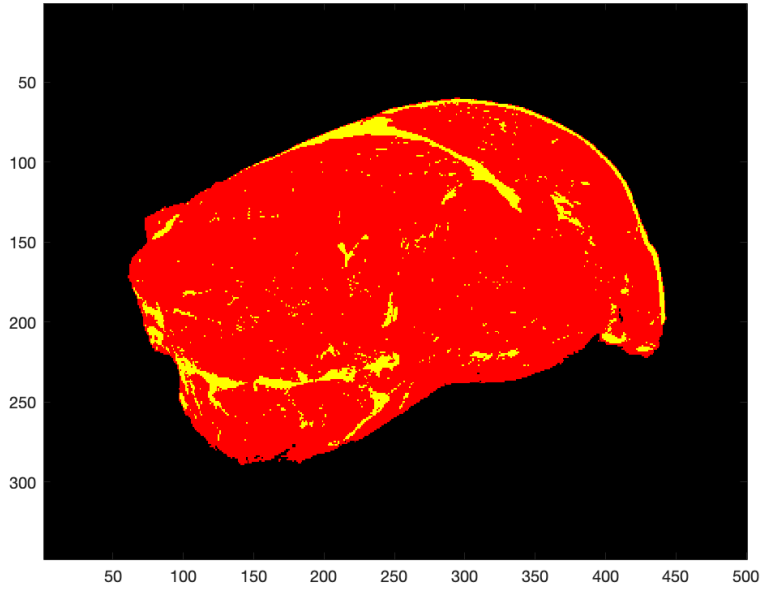
Dall'istogramma degli scores notiamo 3 gruppi:

1. I valori negativi rappresentano lo sfondo;
2. I valori compresi tra 40 e 190, rappresentano la carne;
3. I valori superiori a 190, rappresentano il grasso.



# PCA esplorativa – Soglie di valori (2)

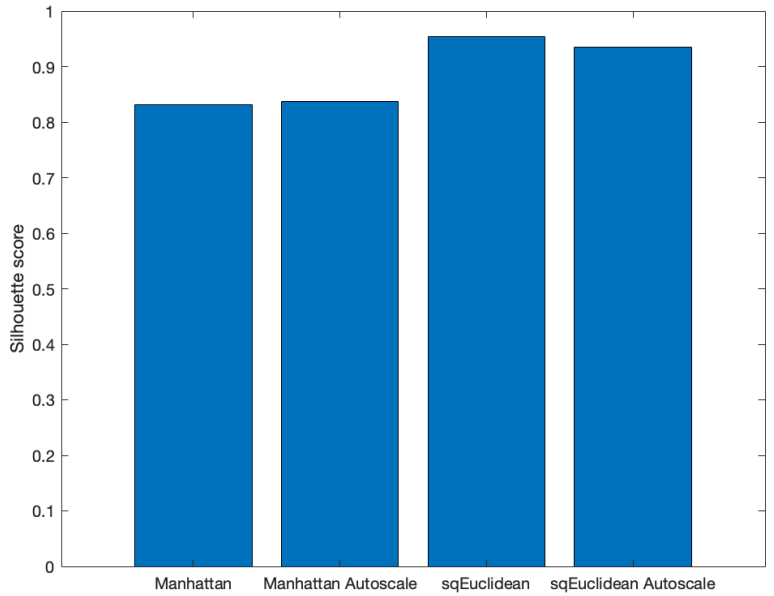
Abbiamo colorato sfondo, carne e grasso rispettivamente di nero, rosso e giallo secondo i valori di soglia selezionati.



# Clustering – KMeans (1)

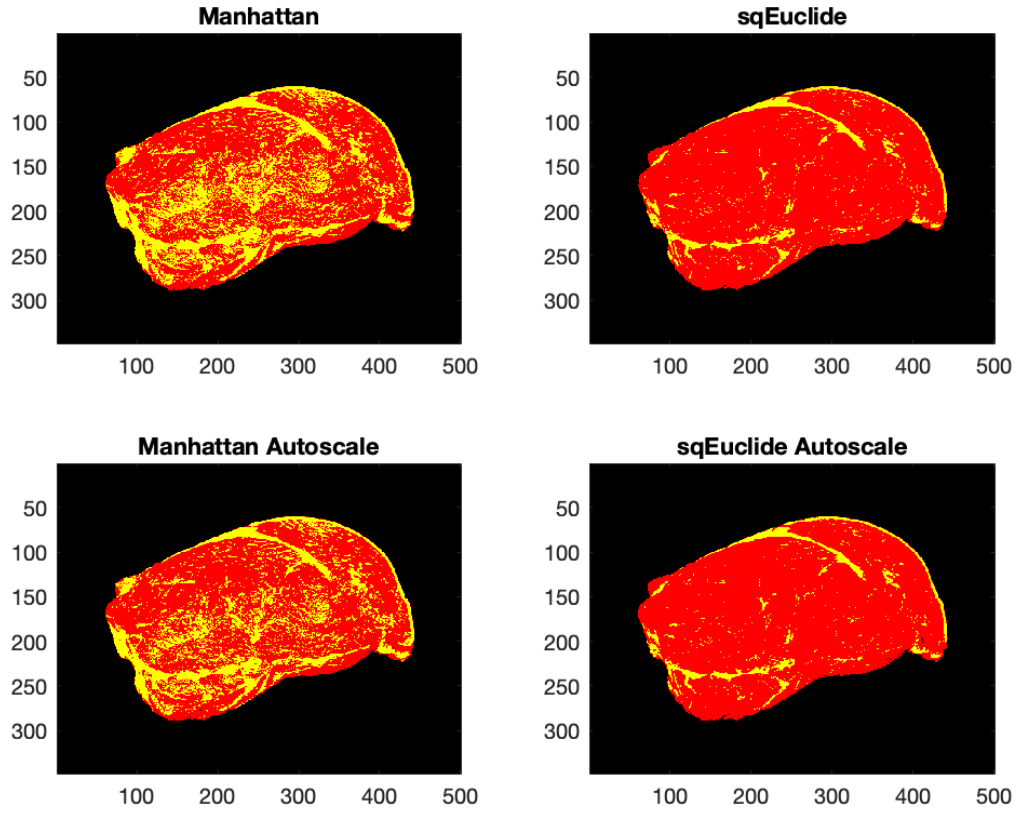
Tramite i silhoutte scores abbiamo confrontato due tipi di distance (Manhattan e Euclidea) con o senza autoscaling.

Abbiamo quindi optato per distanza euclidea **senza** autoscaling.

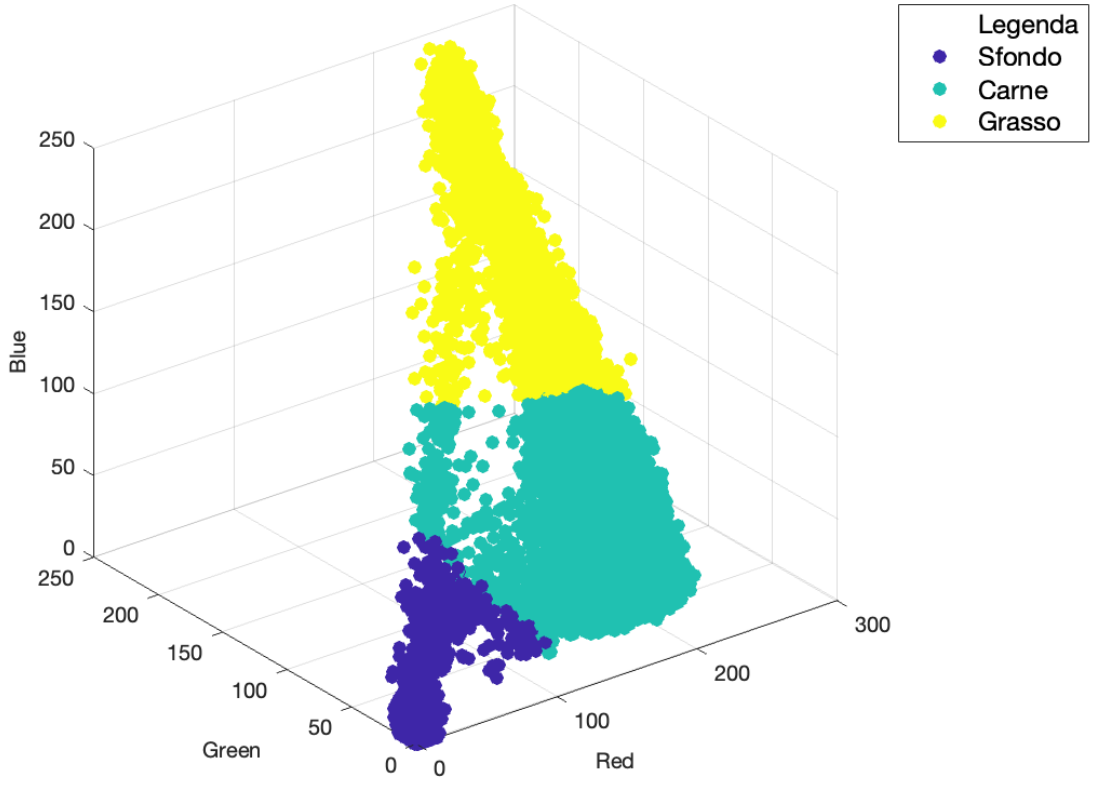


# Clustering – KMeans (2)

Abbiamo colorato le immagini secondo i vari cluster ritornati da KMeans, che hanno confermato quanto già visto con i silhoutte scores.



# Clustering – KMeans (3)

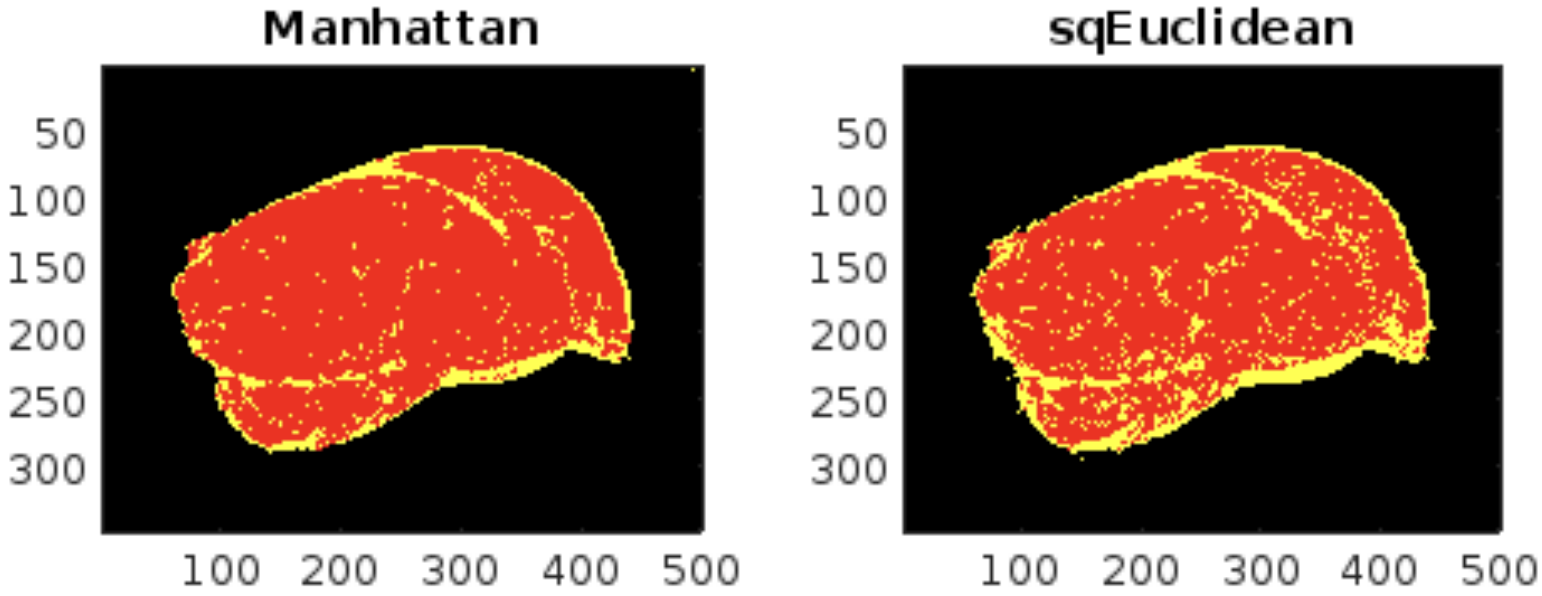



1) Scatter3D KMeans con distanza euclidea

# Clustering – DBSCAN (1)

Lo stesso è stato fatto per DBSCAN impostando i parametri *eps* e *minpts* ottenuti tramite una grid search.

Tuttavia otteniamo dei risultati peggiori, quindi abbiamo scelto di procedere con KMeans.

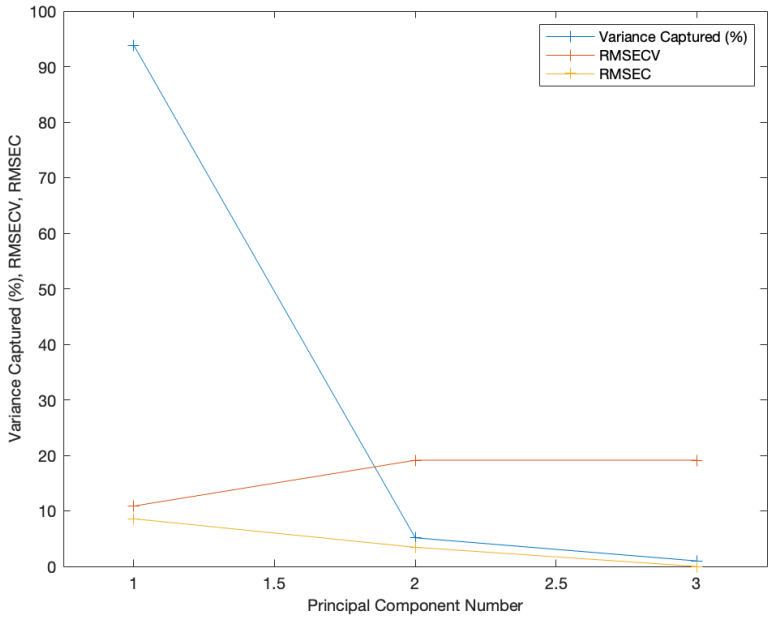


# Clustering - gerarchico

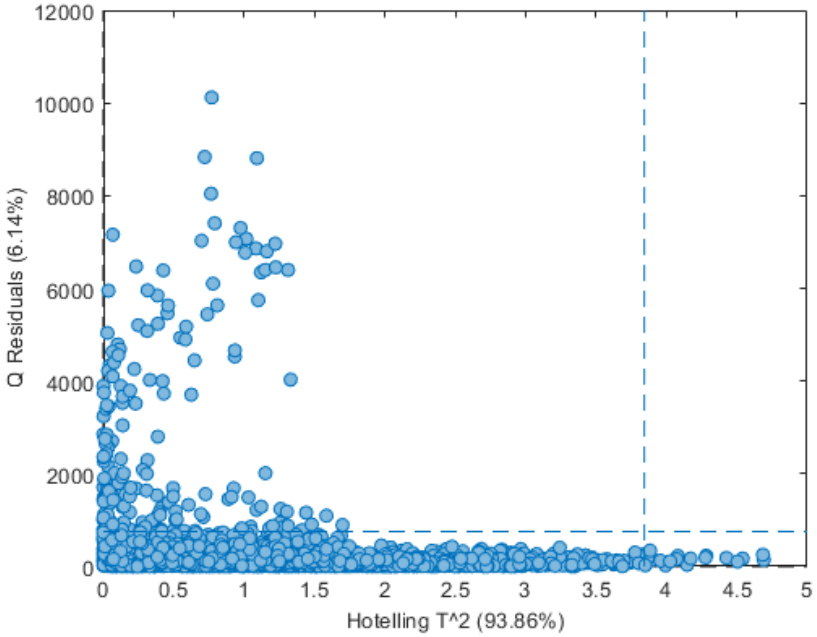
Date le dimensioni del dataset (N = 174000) non è stato possibile eseguire l'algoritmo agglomerativo in quanto ha una complessità pari a $O(n^3)$ e necessita di un quantità di memoria pari a 225 GB.

# Applicazione del modello - Modello sul grasso(1)

Tramite KMeans abbiamo selezionato i pixel relativi al grasso
e abbiamo costruito un modello PCA ad una componente.

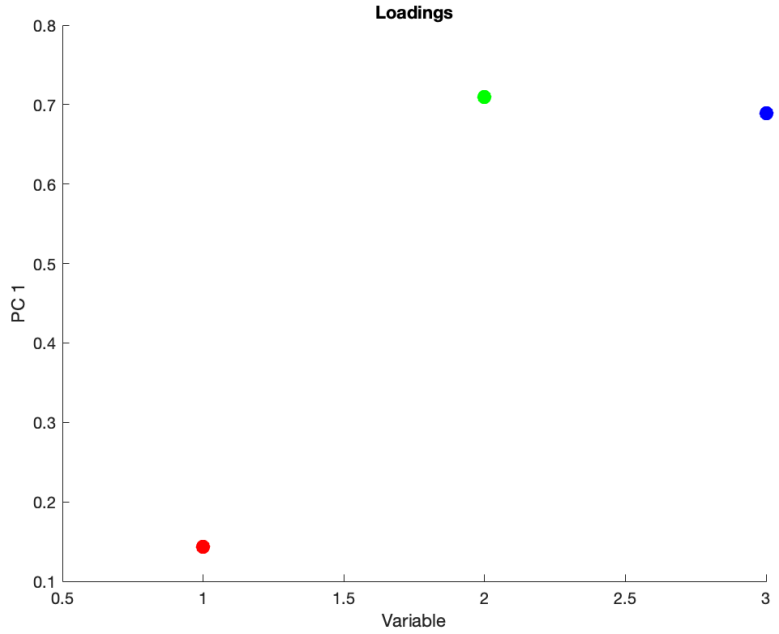



1) RMSEC, RMSECV e varianza delle 3 PC

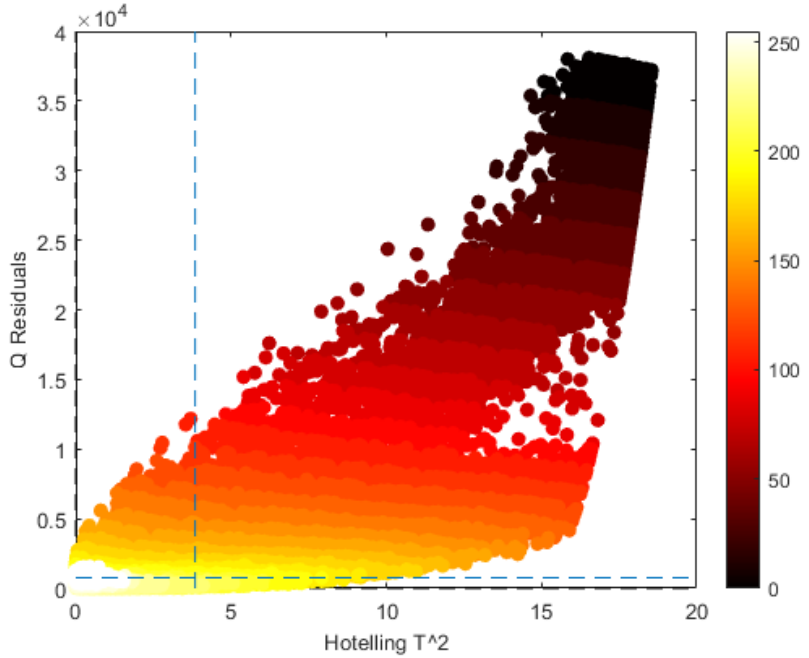



2) Hotelling $T^2$ e Q Residual

# Applicazione del modello - Modello sul grasso(2)

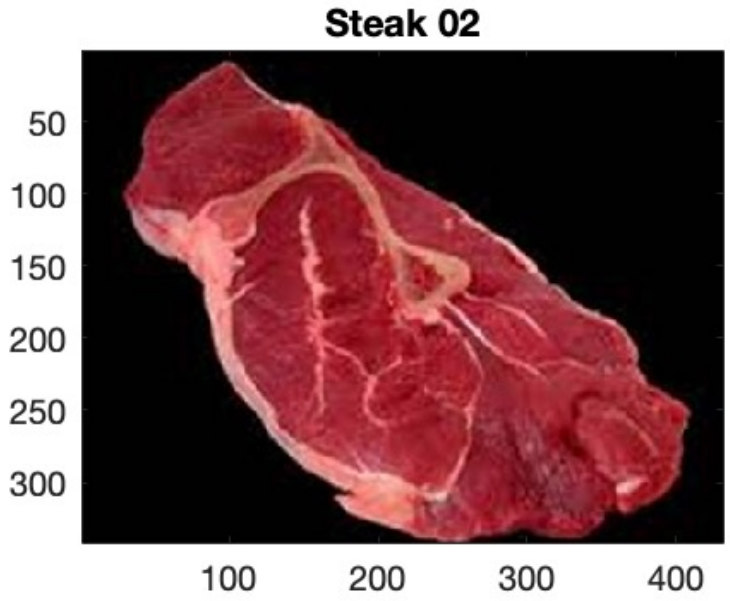


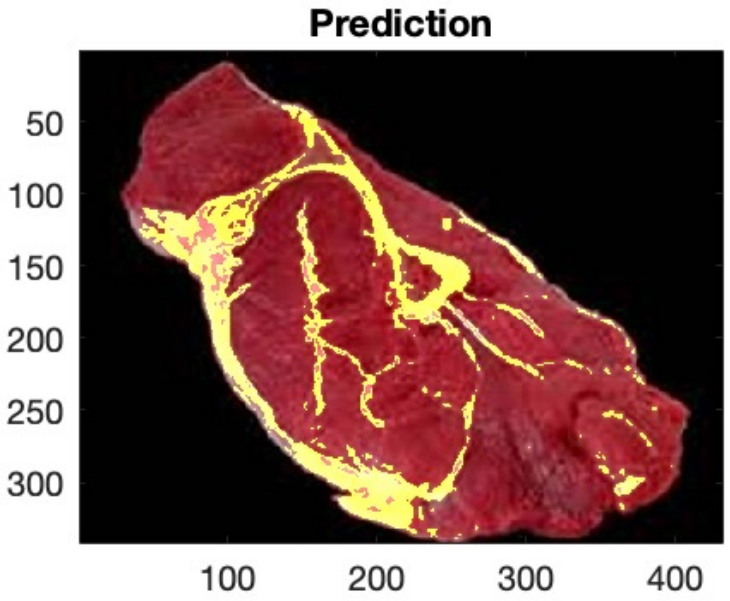

1) Loadings del modello con solo grasso




2) Hotelling $T^2$ e Q Residual dei pixel della seconda immagine proiettati nello spazio del modello definito dal solo grasso

# Applicazione del modello - Risultato




1) Immagine originale

2) In giallo i pixel accettati dal modello costruito sul grasso