Pedol Enrico, Colapinto Giulia, Quadrelli Emanuele

Migliorati Sonia

Statistica Computazionale

10 Gennaio 2026

CRANIAL CLUSTERS

INTRODUZIONE

Il progetto svolto in queste settimane si occupa, come evoca il nome, di cercare eventuali sottogruppi di popolazioni all'interno dei dati.

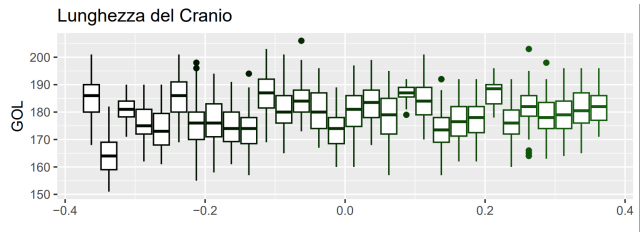
Il dataset analizzato comprende 82 variabili per ciascuna unità statistica, corrispondenti a specifiche misurazioni craniometriche degli individui osservati.

Nel training-set, inoltre, sono note il sesso e la popolazione di appartenenza di ogni cranio.

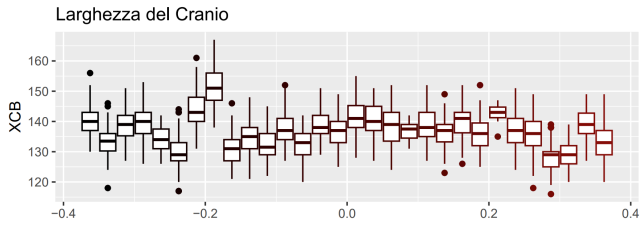
Nel test-set sono indicate anche altre variabili poco utili al fine dell'analisi, come per esempio la zona di ritrovamento e qualche commento.

ANALISI DESCRITTIVA

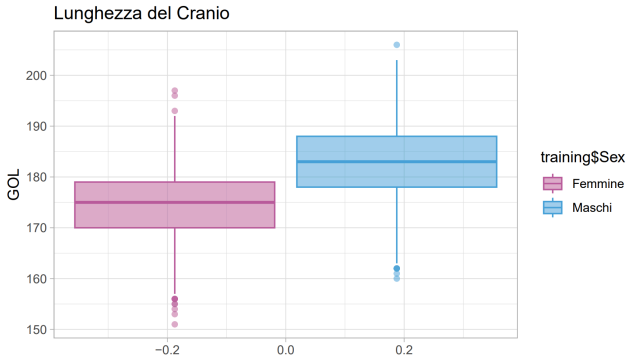
Risulta immediatamente evidente he le due variabili non numeriche del dataset, 'Sex' e 'Population', possano incidere sulle misurazioni.



Come si può notare da questi grafici la popolazione incide notevolmente sulle misurazioni, come la lunghezza del cranio (GOL), la larghezza del cranio (XCB) e non solo.



Riuscire però a individuare 28 gruppi con solamente 2500 osservazioni risulta complesso, l'analisi si è quindi concentrata sul cercare le differenze tra i sessi.



In tutte le principali misurazioni risulta evidente la differenza di grandezza tra i crani maschili e quelli femminili, come si nota dal box plot in figura.

PULIZIA DEI DATI

Prima di tutto sono state eliminate variabili con una grossa quantità di osservazioni pari a 0 (come BSA, SLA, ...). Questo probabilmente poiché spesso nelle suture dei crani di anziani avviene la sinostosi: le suture si fondono completamente e l'osso diventa liscio, questo rende impossibile la misurazione.

I dati sono poi stati centrati, non globalmente usando tutte le osservazioni, ma raggruppandole per popolazione di appartenenza, affinché le caratteristiche potenzialmente peculiari di ogni popolazione non si perdessero.

Per le diverse analisi, ovvero classificazione supervisionata e non, sono state successivamente selezionate diverse combinazioni, per sondare più metodologie e verificare se fosse fondamentale la selezione delle variabili, o se con diverse combinazioni si potessero ottenere risultati pressoché identici.

CLASSIFICAZIONE SUPERVISIONATA (DISCRIMINANT ANALYSIS)

L'analisi è stata condotta separatamente sui set di training e test, previa selezione e centratura dei dati. Il modello di classificazione scelto è stato: **V-Fold Cross Validation**.

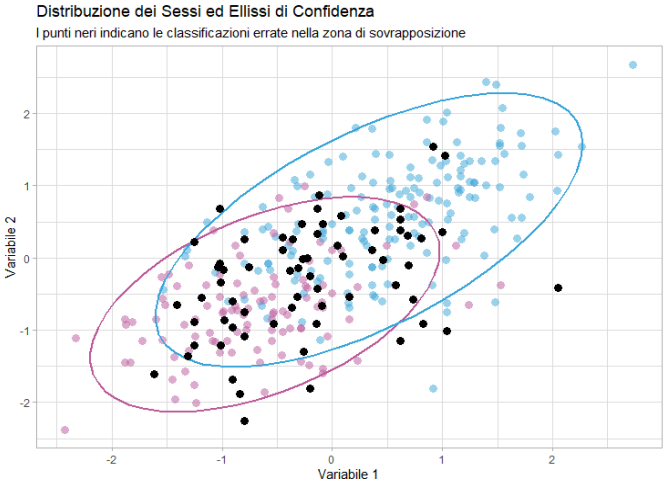
V-FOLD CROSS VALIDATION CON 18 VARIABILI

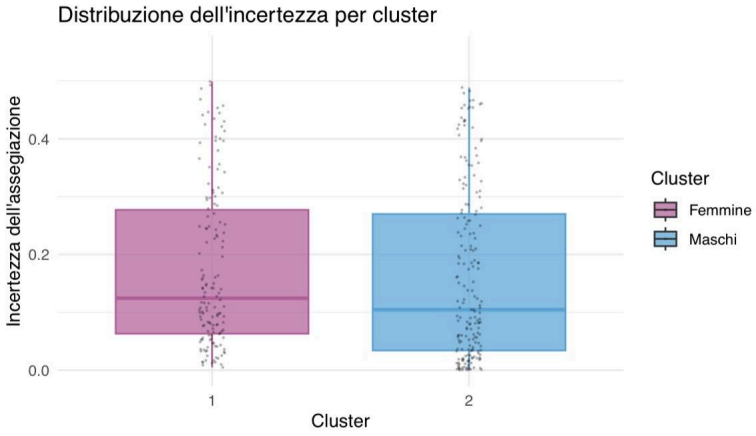
Il primo passaggio compiuto è stata la selezione delle variabili: attraverso l'analisi dei loadings solo sulla prima componente principale. Sono state identificate **18 variabili** chiave che contribuiscono maggiormente alla varianza del dataset (tra cui spiccano NAR, GOL, JUB, ZYB e BNL). Questa selezione ha permesso di ridurre la dimensionalità senza perdere grosse quantità di informazioni e mantenere l'efficacia del modello.

La scelta di testare diverse segmentazioni (**V = 5, 8, 10, 12, 15**) serve a mappare la stabilità del dimorfismo sessuale nel campione. Dividendo il dataset training in V blocchi, lo script utilizza a rotazione V-1 blocchi per l'addestramento e il blocco rimanente per la validazione. Durante tutto il processo di stima dei modelli il valore dell'accuracy si aggira **tra l'80% e l'83%** nel test set, con valori di CV attorno a 0,125. I massimi di accuratezza del modello in media sono stati trovati con una divisione in 10 blocchi. All'interno dello script sono stati valutati solo modelli con proporzioni libere dato l'adattamento ai dati migliore. Lo script evidenzia una competizione tra tre modelli principali di mistura gaussiana:

- Gaussian_pk_Lk_C (**VEE**) e Gaussian_pk_L_D_Ak_D (**EVE**): Modelli migliori nella massimizzazione dell'accuracy del test set;

- Gaussian_pk_L_C (**EEE**): Modello migliore nella minimizzazione del CV del training test.





V-FOLD CROSS VALIDATION CON PCA SCORES (PRIME DUE COMPONENTI)

In questa fase è stato usato un procedimento simile all'analisi precedente ma utilizzando gli **scores delle prime due componenti principali**. Nonostante la drastica riduzione alle due componenti principali (PC1 e PC2), l'accuracy della classificazione nel test set rimane stabile intorno all'82%.

Per verificare la robustezza del modello ed evitare che i risultati siano condizionati da una particolare suddivisione dei dati, il processo di V-Fold Cross Validation è stato fatto con un numero variabile di blocchi: **10, 20, 60, 100**.

I risultati appaiono costanti indipendentemente dal numero di partizioni impostate: l'accuratezza della previsione nel test set non subisce fluttuazioni significative, stabilizzandosi attorno all'**80-83%**. Mentre per il CV del training set il valore di media è **0,145**. I modelli predominanti sono:

- Gaussian_pk_L_D_Ak_D (**EVE**): Appare con maggiore frequenza, specialmente nelle partizioni da 10 e 20 blocchi, in linea con i risultati del modello precedente;
- Gaussian_pk_L_C (**EEE**): Presente in modo significativo nelle partizione di 60 blocchi;
- Gaussian_pk_Lk_Dk_A_Dk (EEV): Sebbene meno frequente come "miglior modello per criterio CV", questo modello emerge spesso quando si cerca massima accuratezza nelle partizioni a 60 e 100 blocchi.

<u>In conclusione</u>, in entrambi i metodi utilizzati i risultati risultano essere molto simili:

- Il modello prescelto è *Gaussian_pk_L_D_Ak_D (EVE)* o in alternativa *Gaussian_pk_Lk_C (VEE)*;
- L'*accuracy* è attorno all'*82%*;
- Il *CV* è tra *0.12 e 0.15*.

Il modello ottenuto può essere considerato sufficientemente accurato nella previsione dei dati del test set.

CLUSTERING NON SUPERVISIONATO (K-MEANS)

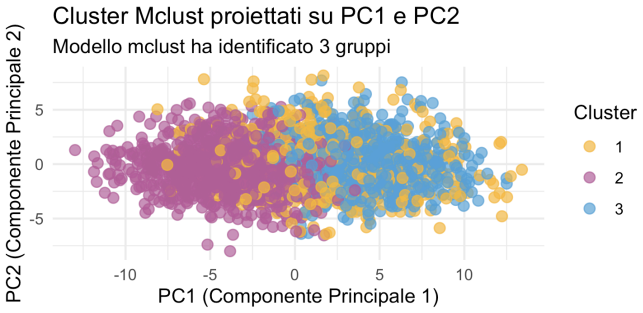
Il K-means clustering risulta poco funzionante con un Adjusted Rand Index (ARI) pari a 0.32 e un errore pari al 21%. Questo è causato dalla sovrapposizione rilevante dei due gruppi la quale conferma come l'origine geografica o la variabilità genetica individuale possano generare somiglianze tra soggetti di sesso diverso, complicando la separazione tra i due gruppi.

MCLUST CON SELEZIONE TRAMITE PCA

Ancora una volta, per la scelta delle variabili su cui applicare il modello di *clustering,* sono state utilizzate le componenti principali. Si è deciso di selezionare le dieci variabili con i pesi più alti nella prima componente principale (quelle con *loadings* maggiori di 0,15).

Una volta selezionate, le variabili sono state inserite in un tibble utilizzato nella funzione *mclust* senza imporre vincoli di alcun genere. Il modello individuato in assenza di restrizioni classifica le osservazioni in tre differenti gruppi, per i quali tuttavia non si riesce a riscontrare una chiara caratteristica discriminante. Osservando le unità statistiche, si ipotizza che il modello identifichi correttamente i crani appartenenti al genere femminile e maschile; tuttavia, esso appare incerto sulle osservazioni estreme dei due sessi, le quali non vengono classificate in nessuno dei due gruppi principali.

Preso atto di ciò, si è deciso di imporre a *mclust* il limite di due gruppi. Il risultato è che il modello migliore è "VEE" (*ellipsoidal, equal shape and orientation*), il quale presenta una bassa entropia; tramite una breve simulazione Monte Carlo è stato riscontrato un *error rate* compreso tra il 13.5% e il 14%. Sebbene i risultati siano notevoli, il procedimento utilizzato non è stato ritenuto soddisfacente, dato che senza l'imposizione di vincoli il modello ottimale riscontra tre gruppi non identificabili.



MCLUST CON SELEZIONE TRAMITE STOCHASTIC FEATURE SELECTION

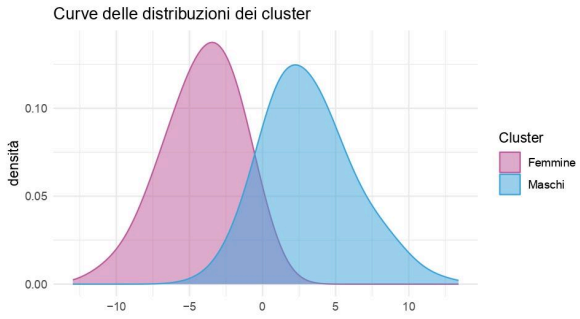
Come soluzione al problema riscontrato, si è deciso di applicare un metodo di selezione delle variabili casuale, denominato *Stochastic feature subset selection.* Ancora una volta sono state considerate le variabili con i *loadings* più alti della prima componente principale. E' stato deciso di selezionare inizialmente venti variabili ed escludere quelle altamente correlate con quelle già presenti.

Definito l'insieme di partenza, è stata costruita una simulazione che estraesse casualmente dieci delle variabili iniziali e, con quel campione, eseguisse sei volte la funzione *mclust,* archiviando in una lista il nome delle variabili utilizzate, il migliore modello individuato e la media del relativo *error rate* su sei simulazioni.

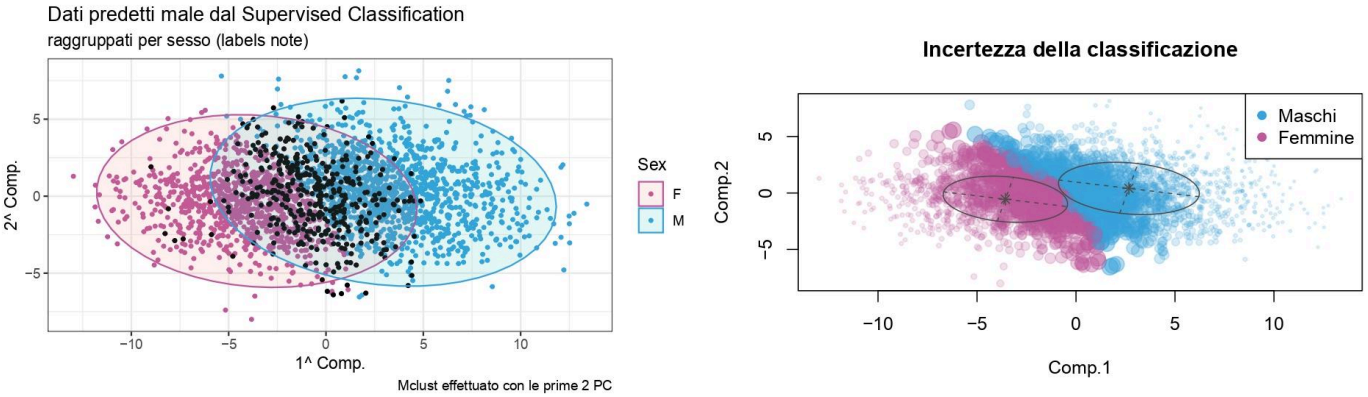
Il risultato ha indicato come modello migliore il "*VEE*", ottenuto utilizzando le variabili "*MAB", "FRC", "AUB", "GOL", "NOL", "AVR", "JUB", "ZYB", "NPH"* e *"FMB"*. Questa volta il modello trovato riconosce i due gruppi senza bisogno di imporre ulteriori vincoli. Svolgendo un'ulteriore verifica attraverso il metodo Monte Carlo si evince che l'*error rate* è compreso tra il 12.9% e il 13.4%.

Ulteriori conferme derivano dal calcolo dell'accuratezza attorno all'87% e dagli indicatori di bontà del modello *R^2 Determinante e R^2 Traccia* che risultano rispettivamente 0.31 e 0.57. Infine, è stata calcolata la divergenza di *Kullback Leibler* il cui valore di 5.84 è risultato molto soddisfacente.

Infine, il modello è stato applicato per prevedere nuove osservazioni tratte dal set di dati di *Test.* In seguito all'omologazione delle variabili con quelle utilizzate nel modello, è stata utilizzata la funzione predict. I risultati sono soddisfacenti dato che, su 303 osservazioni, il modello ne prevede correttamente 252 con un accuracy del 0.83%.



In definitiva, il modello ottenuto interpreta bene le differenze biologiche tra i crani di genere maschile e femminile, con degli errori fisiologici dovuti alla variabilità morfologica individuale naturale presente negli individui, e dimostra un'ottima capacità predittiva su nuove osservazioni.



Opere citate

Howells, W.W. (1973) Cranial variation in man: A study by multivariate analysis of patterns difference among recent human populations.

Cambridge, MA: Peabody Museum of Archeology and Ethnology, Harvard University.