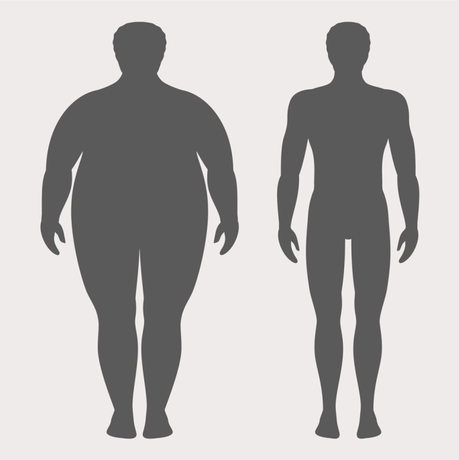
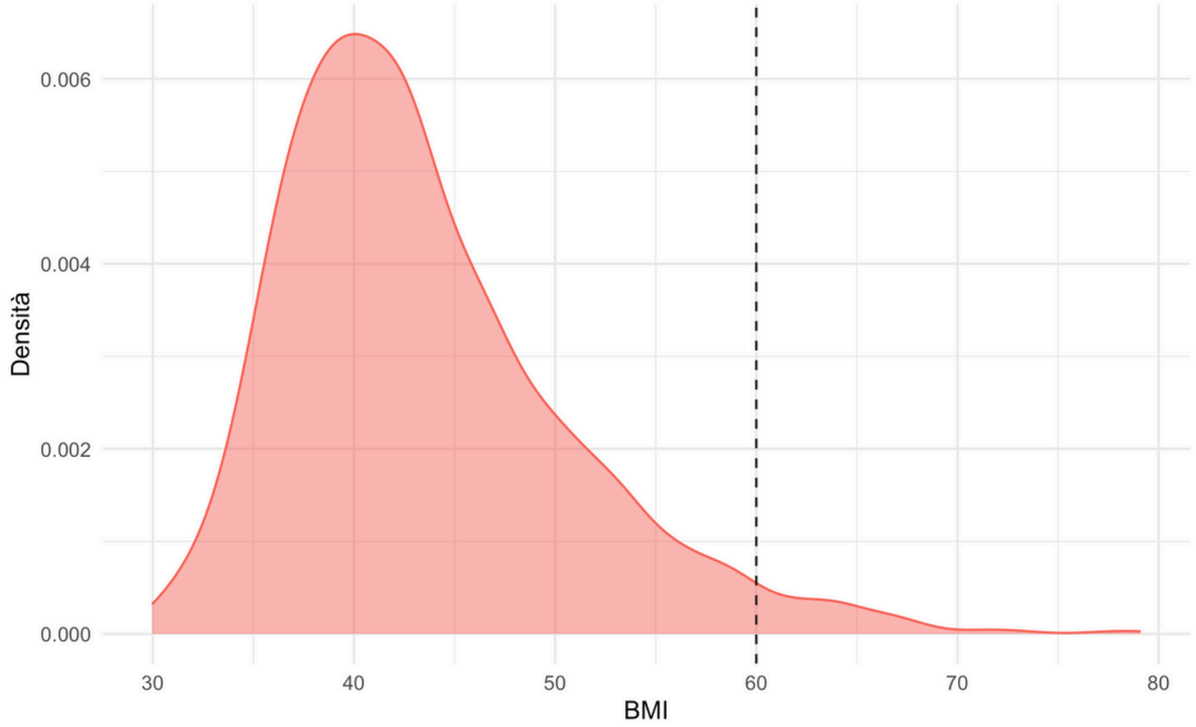
Analisi comparativa tra Stepwise e PCA per l'identificazione delle variabili chiave nella Perdita di Peso.



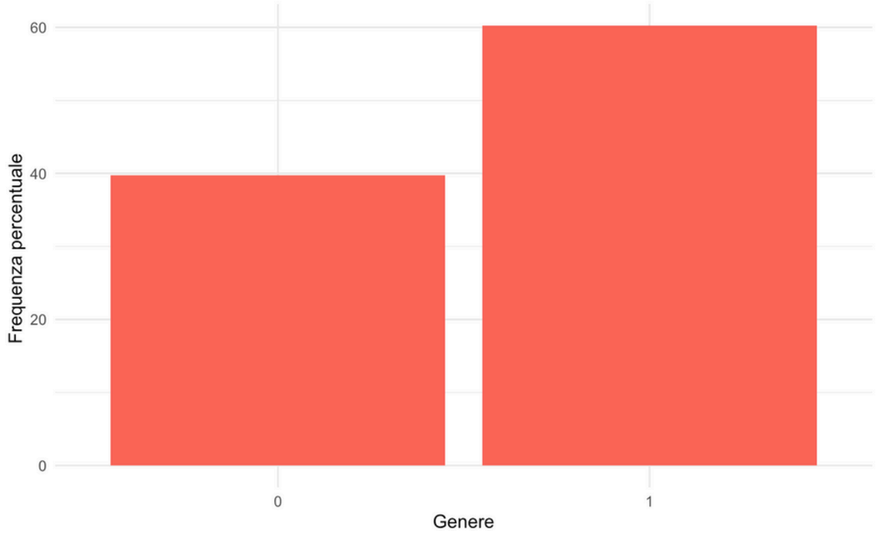
Aniello De Palma
Nicolò Bonato

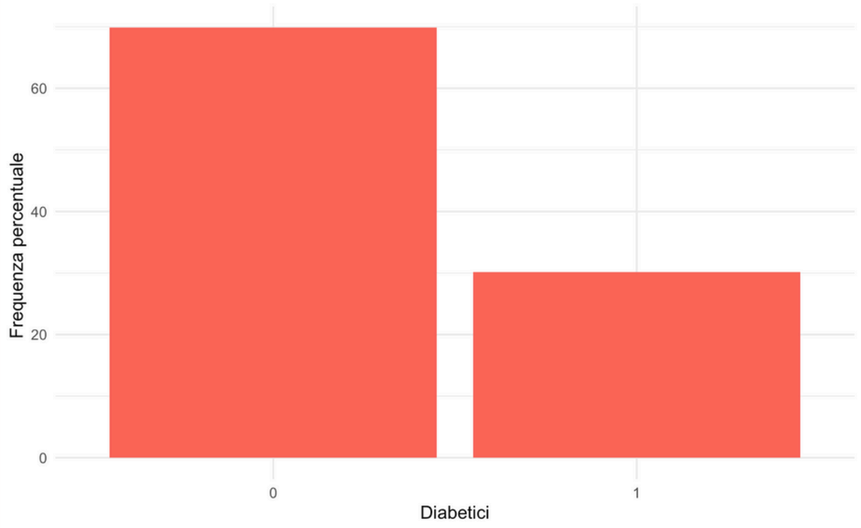
Descrizione dei dati

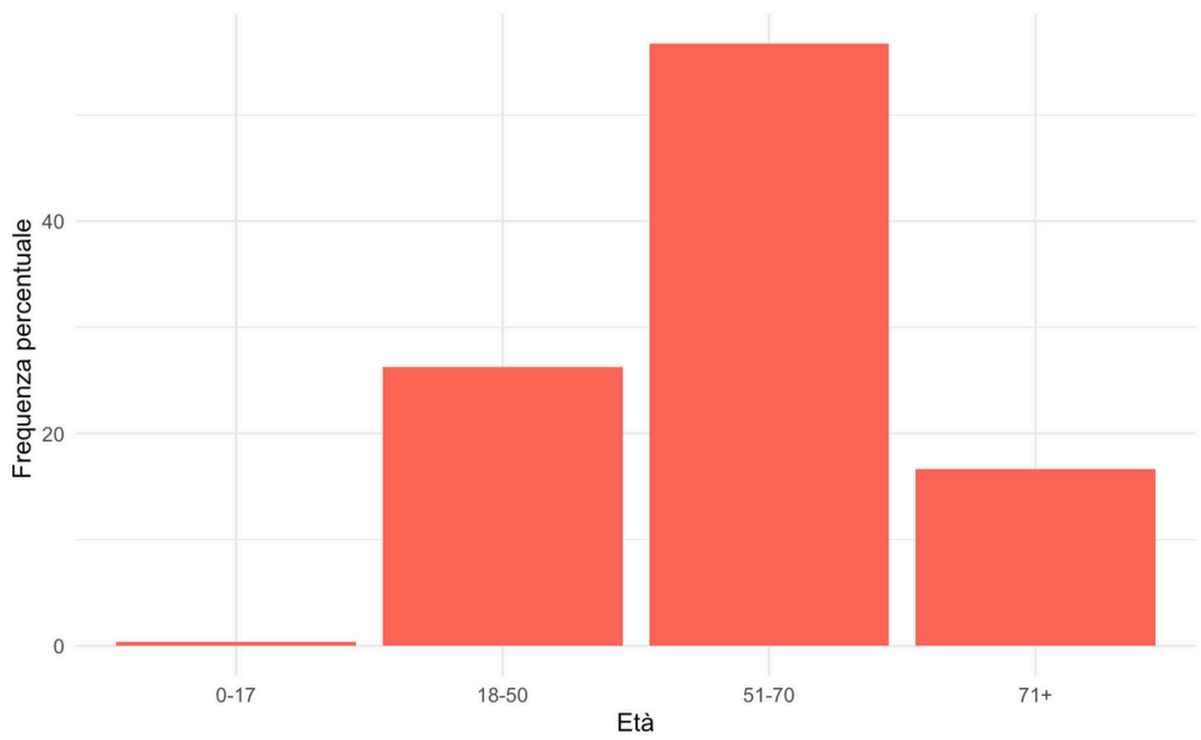
I dati raccolti sono relativi a un campione composto da 2.076 individui affetti da obesità, con un set di 562 variabili registrate che forniscono un quadro dettagliato delle caratteristiche fisiologiche e cliniche dei pazienti.



Alcune caratteristiche dei pazienti







Data cleaning

Step 1

Gestione dei missing

Step 2

Ricodifica delle variabili

Step 3

Eliminazione delle variabili con più del 25% di missing

Step 4

Eliminazione delle variabili ridondanti

Step 5

Imputazione dei missing

Step 6

Creazione della variabile outcome

Variabile outcome

Perdita di peso maggiore del 5%

$$\frac{\text{peso iniziale} - \text{peso finale}}{\text{peso iniziale}}$$

STEPWISE: Procedura

Step 1

Divisione del dataset in train e test set

Step 2

Creazione del modello nullo

Step 3

Creazione del modello completo

Step 4

Indiviuazione del criterio di selezionamento : BIC

Step 5

Selezione delle variabili con il metodo “both”

Step 6

Risultati

STEPWISE: Risultati modello logit

La Stepwise ha selezionato sei variabili statisticamente significative per la variabile outcome:

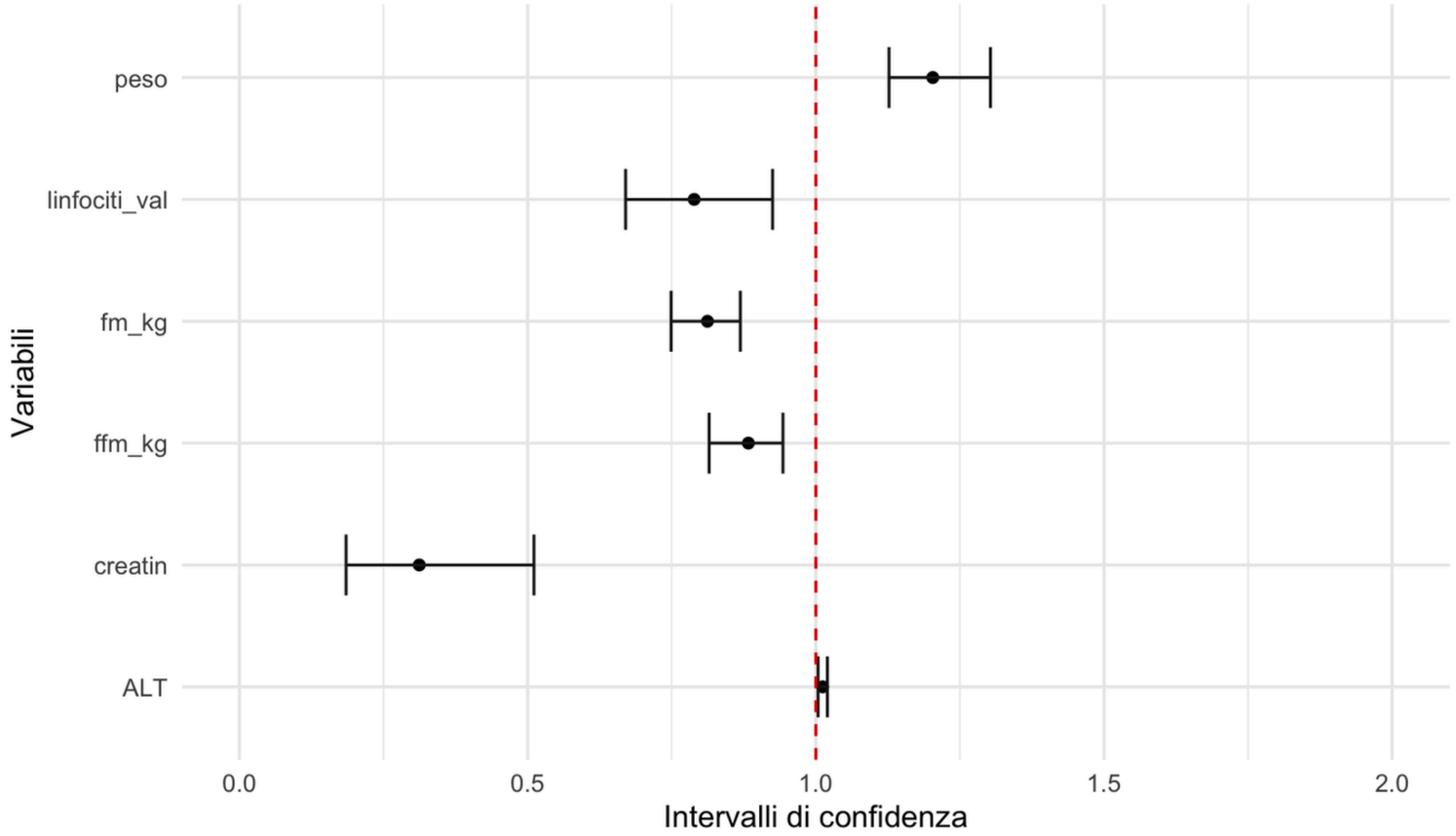
- **ffm_kg** è la massa magra
- **fm_kg** è la massa grassa
- **creatin** è la quantità di creatina nell'organismo
- **peso** è la misurazione della massa corporea iniziale
- **ALT** è il livello di *Alanina Aminotransferasi*, enzima coinvolto nel metabolismo degli aminoacidi, nell'organismo
- **linfociti_val** è il numero di linfociti nel sangue

Variabili	Coefficienti	p-value
Intercetta	-1.502425	0.000365
ffm_kg	-0.124656	0.000751
creatin	-1.165971	6.97e-06
peso	0.184978	5.16e-07
fm_kg	-0.207866	3.80e-08
ALT	0.011652	0.003537
linfociti_val	-0.236776	0.003970

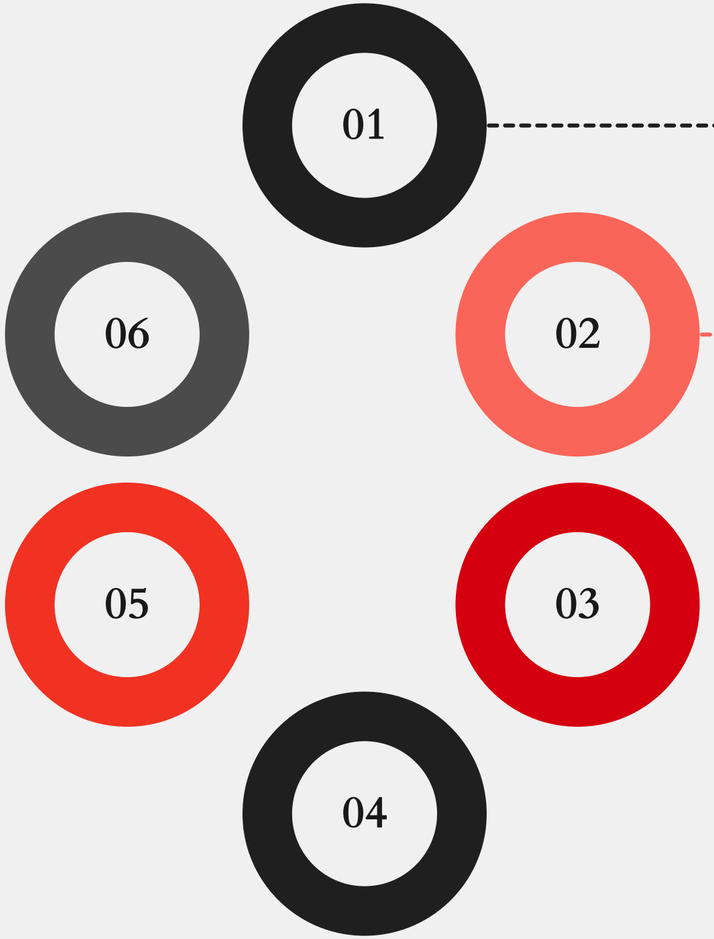
Modello logit con le variabili selezionate attraverso la stepwise

Odds-Ratio

Intercetta	ffm_kg	creatin	peso	fm_kg	ALT	linfociti_val
0.2225898	0.8828001	0.3116199	1.2031918	0.8123159	1.0117199	0.7891679



PCA: Procedura



Step 1

Selezione delle covariate standardizzate

Step 2

Creazione degli scores e dei loadings

Scores e Loadings

Vogliamo trasformare le variabili originarie $X = (X_1, X_2, \ldots, X_{55})^t$ in $Z = (Z_1, Z_2, \ldots, Z_{55})^t$ vettore delle componenti principali.
La trasformazione è di tipo lineare,

$$Z_j = \phi_{1j}X_1 + \phi_{2j}X_2 + \ldots + \phi_{55j}X_{55}$$

viene detta j-esima componente principale. I coefficienti

$$\phi_j = (\phi_{1j}, \phi_{2j}, \ldots, \phi_{55j})^t$$

sono detti *loadings* della j-esima componente.

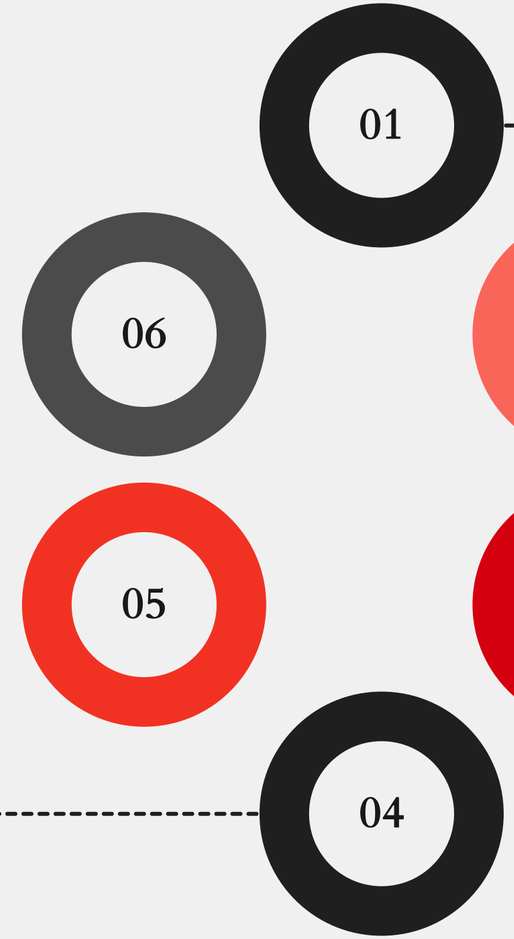
Z_1	Z_2	Z_3	$\ldots$	Z_{54}	Z_{55}
−2.0895	−3.8574	0.2484	$\cdots$	−0.2321	−0.1503
−3.0588	−2.0520	0.1935	$\cdots$	−0.2764	0.0285
−3.0664	−1.2639	−1.0674	$\cdots$	−0.3289	−0.0689
−0.6524	−2.1123	1.8394	$\cdots$	0.0234	−0.0401
−0.6606	1.0846	−0.4657	$\cdots$	0.0541	−0.0063

Table 1: Head Scores

	ϕ_1	ϕ_2	ϕ_3	$\ldots$	ϕ_{54}	ϕ_{55}
eta	0.1123	0.0057	0.0850	$\cdots$	0.0055	$3.27e^{-3}$
qualific.	−0.0622	0.0623	−0.0038	$\cdots$	−0.0015	$-1.09e^{-4}$
job_cat.	0.0312	−0.0315	0.0304	$\cdots$	−0.0006	$7.72e{-5}$
peso	−0.2739	−0.2098	0.1058	$\cdots$	−0.7546	$-2.09e^{-1}$
altezza	−0.2544	0.1054	0.1284	$\cdots$	0.3149	$1.04e^{-1}$

Table 2: Head loadings

PCA: Procedura



Step 1

Selezione delle covariate standardizzate

Step 2

Creazione degli scores e dei loadings

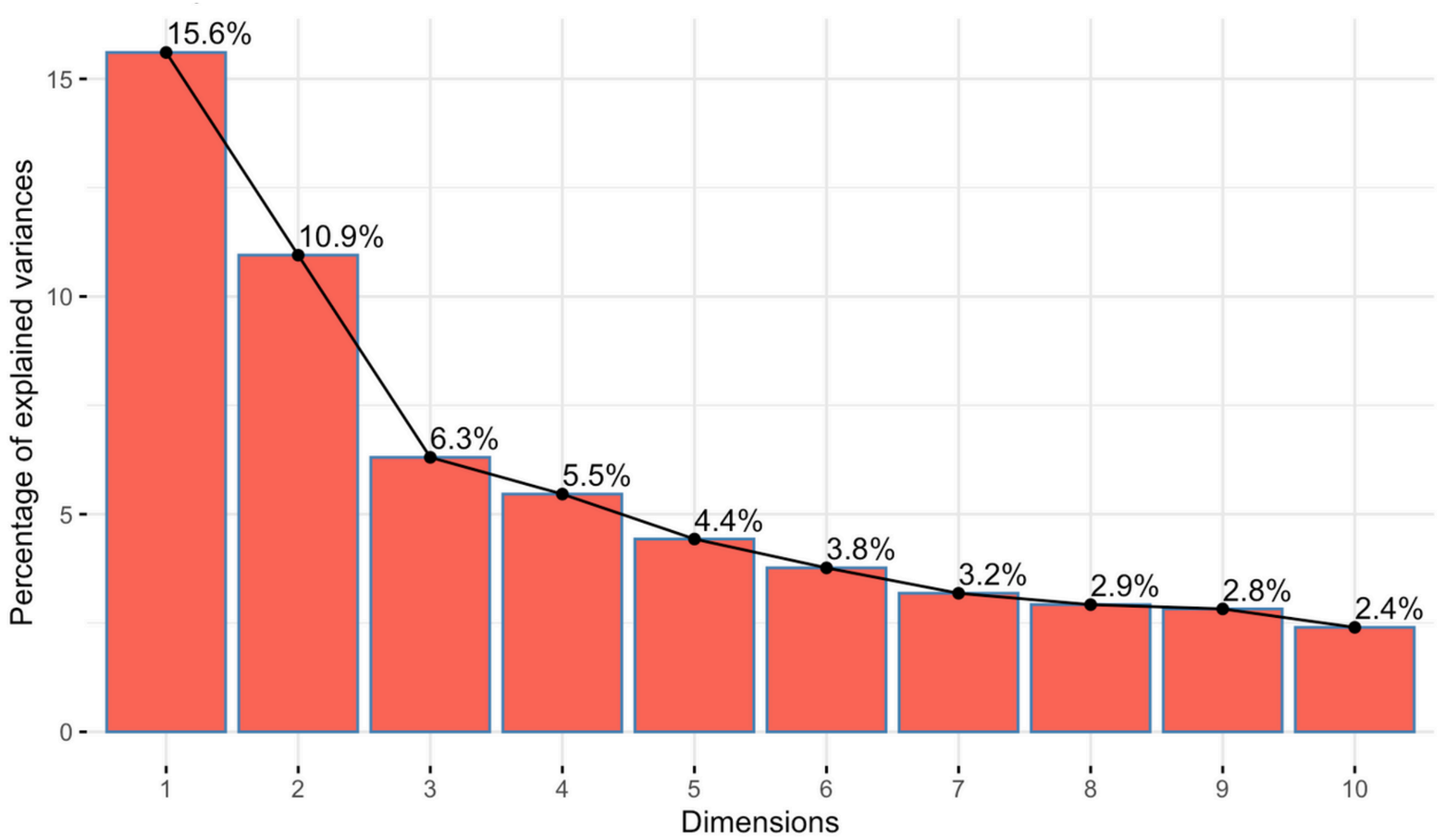
Step 3

Verifica del contributo delle componenti principali alla variabilità totale

Step 4

Screeplot ed elbow method

Screeplot


15.6%
10.9%
6.3%
5.5%
4.4%
3.8%
3.2%
2.9%
2.8%
2.4%
Percentage of explained variances
0
5
10
15
1
2
3
4
5
6
7
8
9
10
Dimensions

PCA: Procedura

Step 1

Selezione delle covariate standardizzate

Step 2

Creazione degli scores e dei loadings

Step 3

Verifica del contributo delle componenti principali alla variabilità totale

Step 4

Screeplot ed elbow method

Step 5

Selezione delle variabili che spiegano maggiormente le PC

Step 6

Verifica empirica

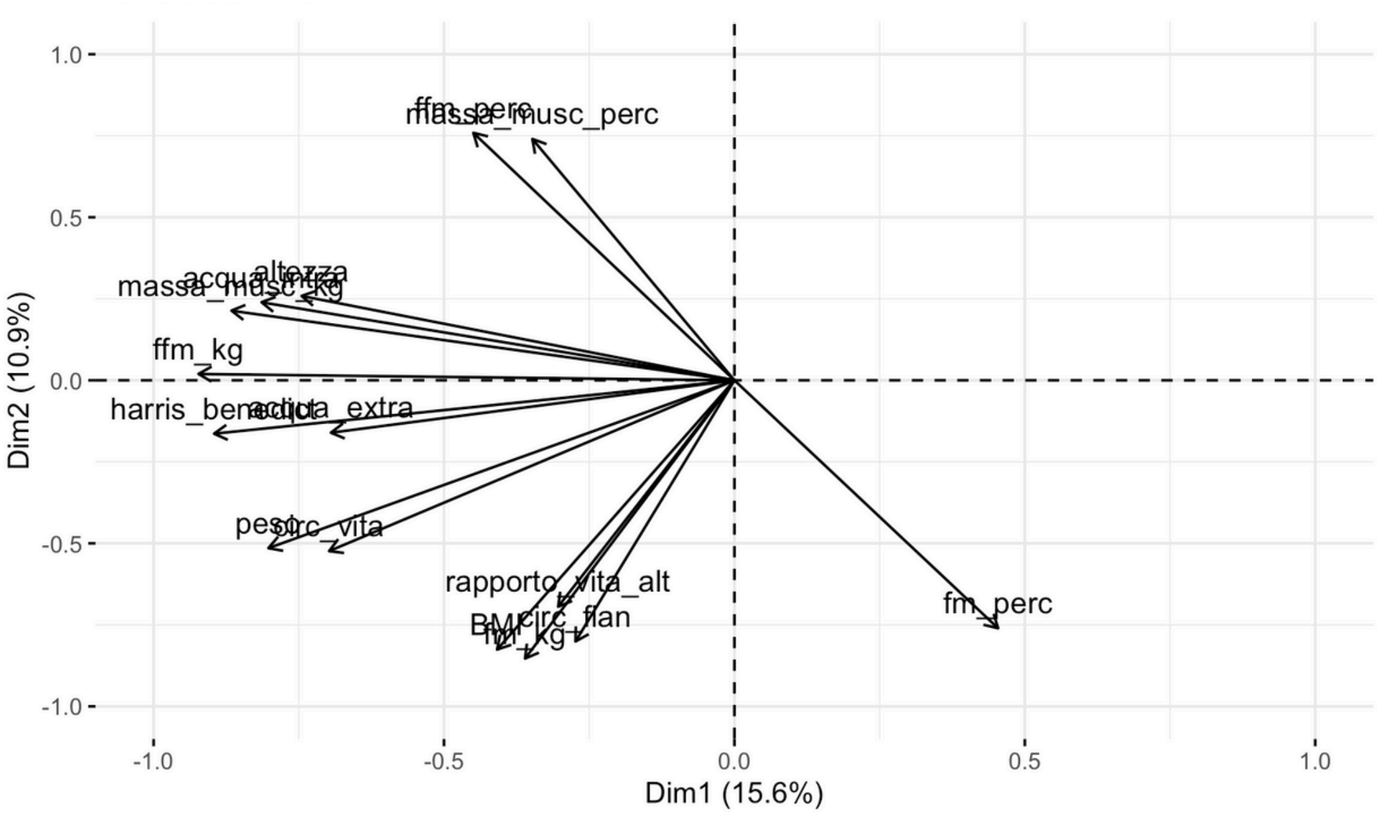
Selezione delle variabili

nome_variaibli	*contributo*
ffm_kg	9.9245
harris_benedict	9.3557
massa_musc_kg	8.7405
acqua_intra	7.7263
peso	7.5038
altezza	6.4722
circ_vita	5.6762
acqua_extra	5.6262

Table 1: Prima PC

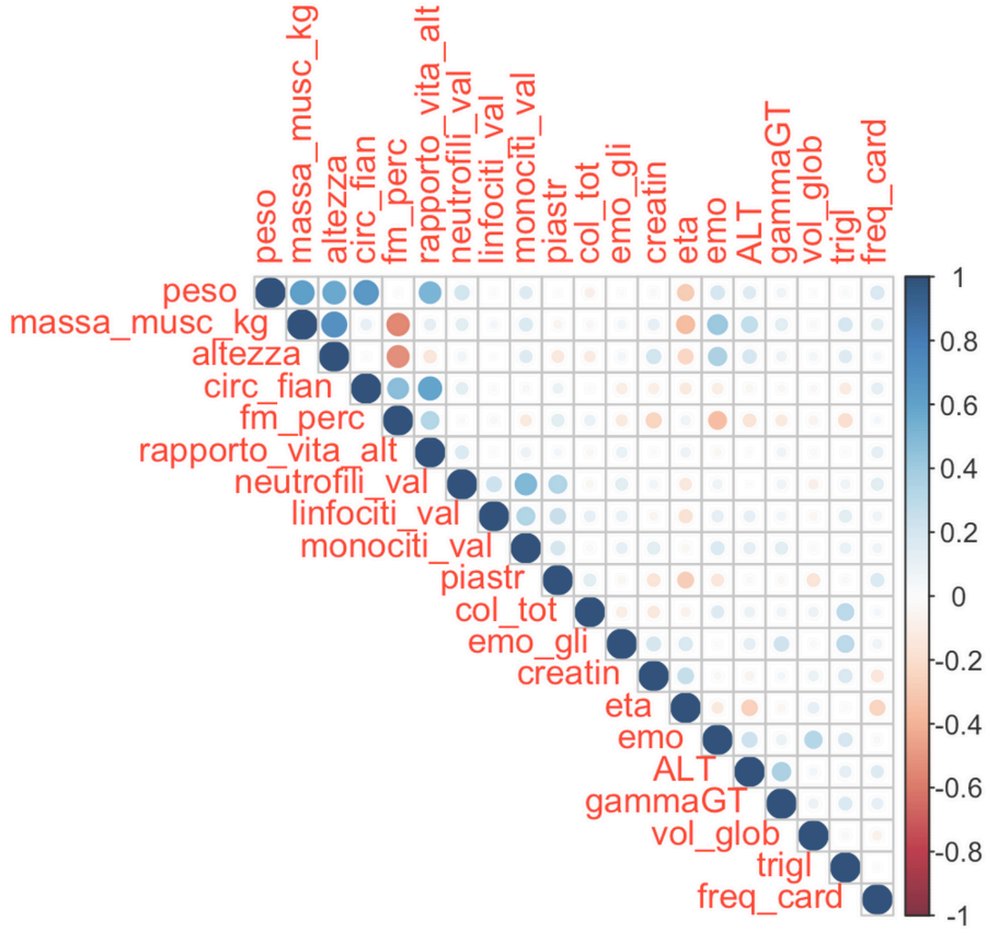
nome_variaibli	*contributo*
fm_kg	12.0971
BMI	11.3285
circ_fian	10.6542
fm_perc	9.6027
ffm_perc	9.5669
massa_musc_perc	9.0990
rapporto_vita_alt	8.0043
circ_vita	4.5583
peso	4.4020

Table 2: Seconda PC



Correlation Plot

Tra le variabili selezionate precedentemente prendiamo in considerazione, per il modello logit, solo quelle con un coefficiente di correlazione minore di 0.7



PCA: Risultati modello logit

Le variabili statisticamente significative sono:

- **peso** è la misurazione della massa corporea iniziale
- **altezza** è la misurazione in cm del paziente
- **fm_perc** è la percentuale di massa grassa
- **neutrofili_val** è il numero di neutrofili nel sangue
- **linfociti_val** è il numero di linfociti nel sangue
- **emo_gli** è il numero di emoglobina glicata, che rispecchia la concentrazione di glucosio nel sangue
- **creatin** è la quantità di creatina nell'organismo
- **ALT** è il livello di Alanina Aminotransferasi, enzima coinvolto nel metabolismo degli aminoacidi, nell'organismo

Variabili	Coefficienti	p-value
Intercetta	8.4270369	0.002042
peso	0.0331009	$1.99e-05$
massa_musc_kg	-0.0059752	0.673689
altezza	-0.0218800	0.099493
circ_fian	-0.0097784	0.207162
fm_perc	-0.1101780	$2.46e-11$
rapporto_vita_alt	-0.1279250	0.914099
neutrofili_val	0.1290235	0.019015
linfociti_val	-0.2416237	0.008235
monociti_val	0.2123979	0.613060
piastr	-0.0015421	0.225802
col_tot	0.0019673	0.267876
emo_gli	-0.0098490	0.040209
creatin	-1.0703689	0.000129
eta	0.0003013	0.959900
emo	-0.0911338	0.101173
ALT	0.0124475	0.009169
gammaGT	0.0005946	0.714021
vol_glob	-0.0059690	0.444436
trigl	0.0012361	0.246335
freq_card	0.0037099	0.495461

Modello logit con le variabili selezionate attraverso la PCA

Confronto dei modelli: Bontà di adattamento

Si parte col confronto dei Pseudo-R^2 . Essi aiutano a selezionare il modello che meglio si adatta ai dati, valutando quello che spiega in modo più efficace le variazioni nella variabile target.

Questa differenza nei Pseudo-R^2 ci invita a considerare ulteriori analisi. Si valuta in questo caso la bontà di adattamento tenendo conto del principio di parsimonia

Pseudo-R^2	Stepwise	PCA
Efron	0.1891388	0.1725211
McFadden	0.1408172	0.1335146
Nagelkerke	0.2348536	0.2237523
Cox-Snell	0.1748825	0.1666160

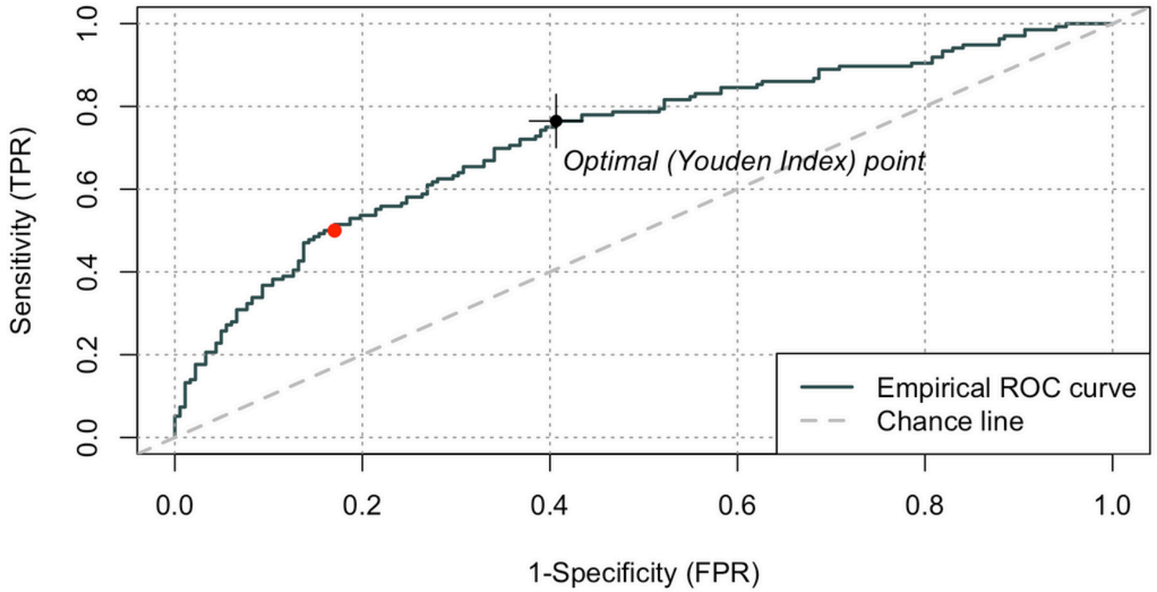
Confronto dei Pseudo-R^2

	Stepwise	PCA
AIC	1507.064	1547.754
BIC	1643.196	1655.886

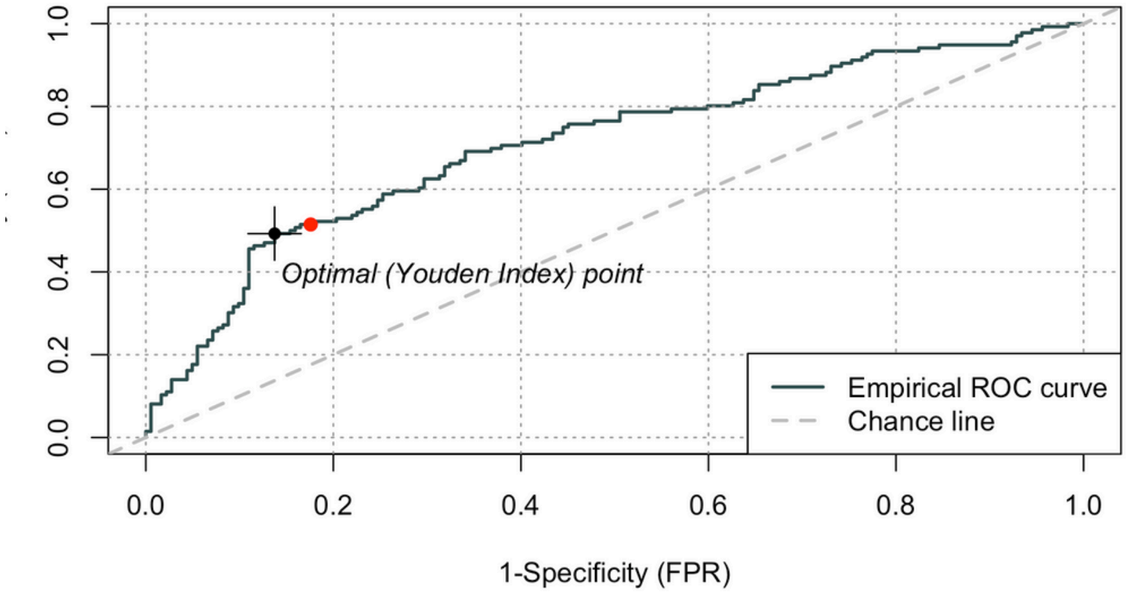
Confronto dei criteri di validazione

Confronto dei modelli: Validazione

Stepwise



PCA



	Stepwise	PCA
Cutoff	0.354	0.518
Accuracy	0.667	0.704
Sensitivity	0.765	0.493
Specificity	0.593	0.863
AUC	0.730	0.710
Test error	0.333	0.296

Confronto delle misure di accuratezza

Conclusioni

- Dopo un’adeguata pulizia dei dati abbiamo confrontato, quindi, due tecniche per la selezione delle variabili significative per la perdita di peso di almeno un 5%.

- La stepwise ha selezionato solo 6 variabili, tutte numeriche continue.

- Tramite l’analisi delle componenti principali, invece, sono state selezionate più variabili, ma non tutte sono statisticamente significative per l’outcome.

- Dal confronto tra i modelli, la Stepwise risulta essere più performante sia in termini di adattabilità ai dati utilizzati per l’addestramento, sia out-sample.

- È importante precisare che la Stepwise è una tecnica tradizionale creata proprio per questo scopo, mentre la PCA è utilizzata principalmente per comprimere le informazioni in meno variabili perdendo, però, parte delle informazioni.