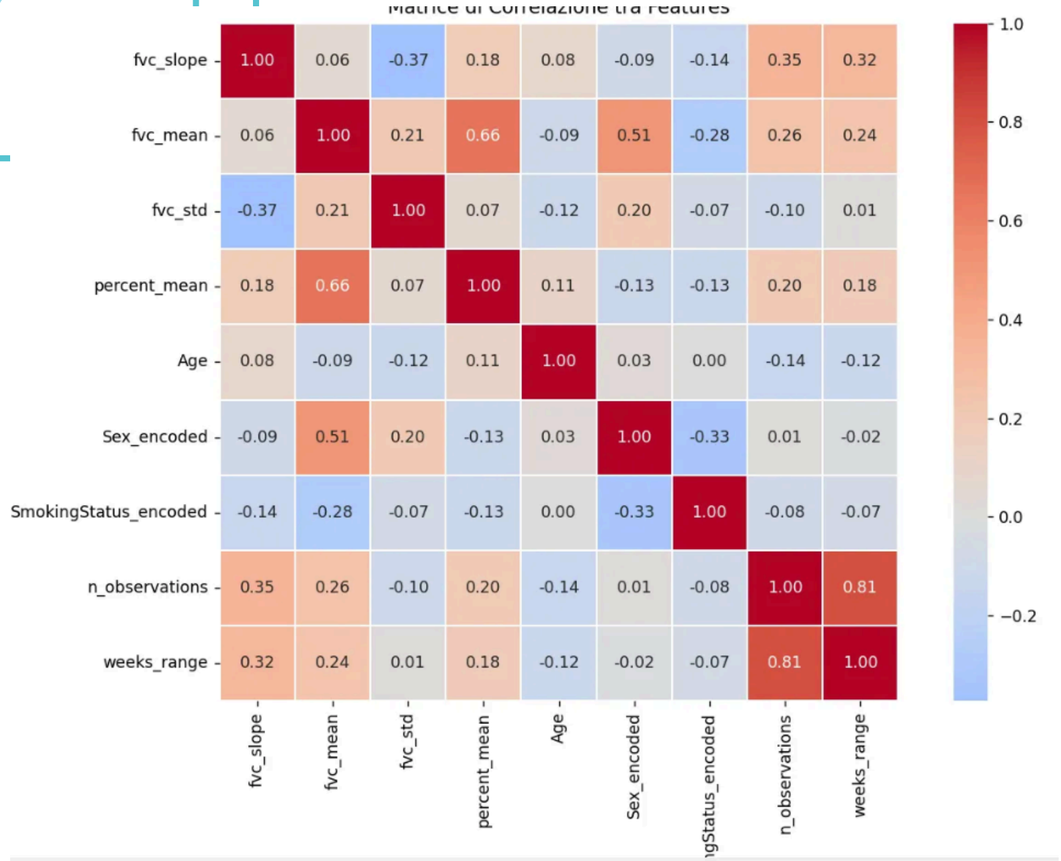
PREDICTION FOR WEEK 0



Clustering

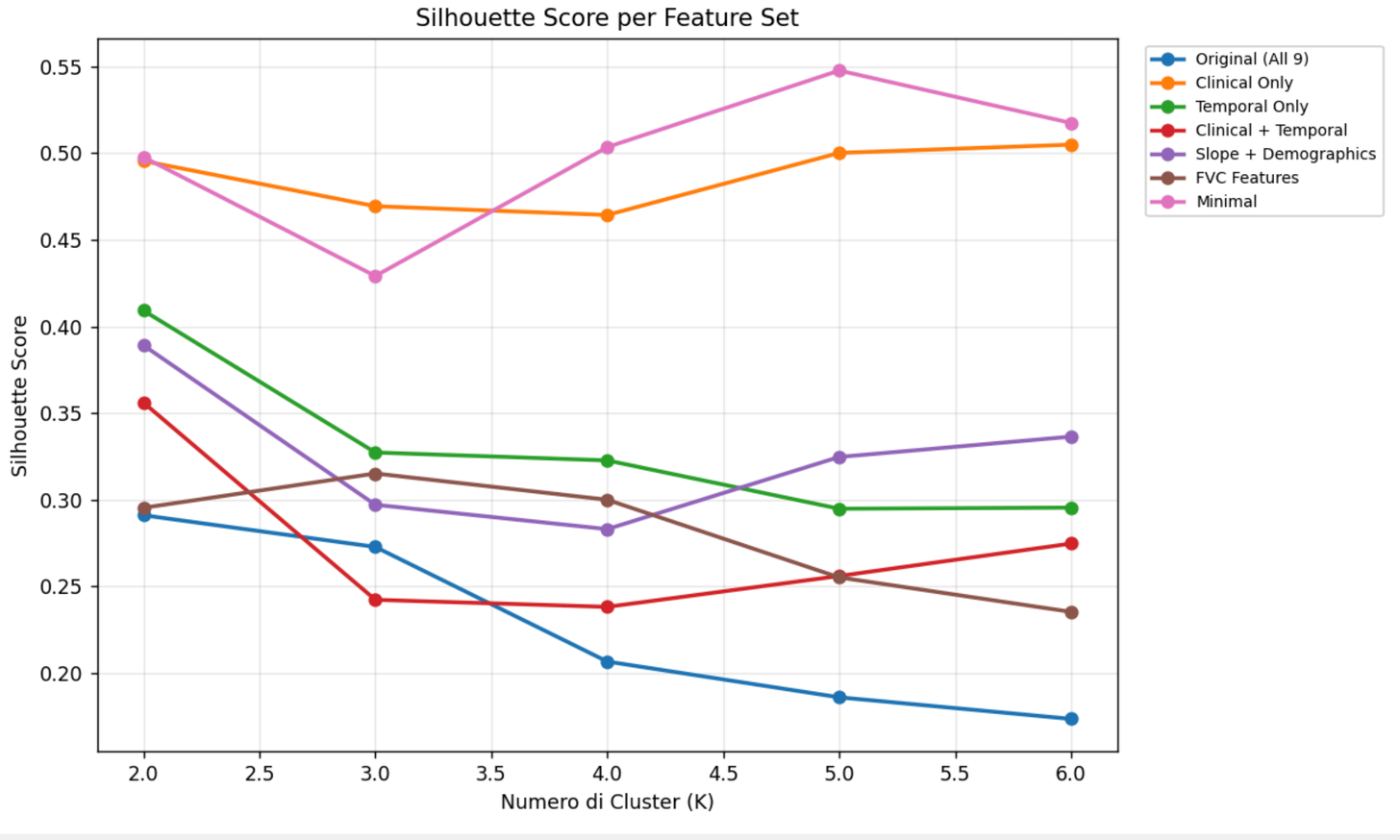
Divide patients into clusters and adapt each linear regression to the cluster to define similar characteristics and get informations from multiple individuals

PREDICTION FOR WEEK 0

Different subsets of features starting from 9 features :

- fvc_slope
- fvc_mean
- fvc_std
- percent_mean
- Age
- Sex_encoded
- SmokingStatus_encoded
- n_observations
- weeks_range

Best one: Minimal (fvc_slope, sex, smokingstatus)



Cluster 0: FVC slope → -1.94

- Only Male
- 80 ex-smoker
- 7 currently smokes
- FVC Decline slow/stable
- MAE 1174.14
- R^2 0.004

Cluster 1: FVC slope → -4.18

- Only Female
- 23 Never smoked
- FVC Decline moderate
- MAE 167.79
- R^2 0.003

Cluster 2: FVC slope → -14.12

- Only Male
- 26 ex-smoker
- 3 never smoked
- Rapid FVC decline
- MAE 456.58
- R^2 0.055

Cluster 3: FVC slope → -3.86

- Only Male
- 23 never smoked
- FVC Decline moderate
- MAE 532.50
- R^2 0.054

Cluster 4: FVC slope → -2.37

- Only Female
- 12 ex-smoker
- 2 currently smokes
- FVC Decline slow/stable
- MAE 68.46
- R^2 0.087

Validation made on the 18 patients with week 0 so the values of MAE and R^2 are very apporoximate.

Comparison between methods:

- **Cluster based on FVC_slope, Sex, Smoking Status**
  - **MAE : 622.04**
- **Individual (based on individual slope → not cluster)**
  - **MAE : 127.62**
- **Cluster based on all 9 features**
  - **MAE : 536.12**

The MAE increases for the first method respect to clustering with 9 features, but the clusters become much more interpretable.

Best approach between the three methods

- **Hybrid Approach**
  - **Use individual method for patients with a good fit,**
  - **Use optimized clusters for patients with variable data , and far from week 0**

Use individual method if MAE% value better than cluster:

- **MAE%< 5 and distance from week 0 < 15 → personal high confidence**
- **MAE%<10 → personal medium confidence**
- **MAE%<12 → personal low confidence**

else use cluster estimate → cluster optimized

With this new method:

- **Personal high confidence (n=87)**
  - **Avg R^2 : 0.450**
  - **Distance week0 : 6.7 weeks**
  - **Avg MAE: 67**
- **Personal medium confidence (n=49)**
  - **Avg R^2 : 0.358**
  - **Distance week0 : 21.1 weeks**
  - **Avg MAE: 98.0**
- **Cluster optimized (n=20)**
  - **Avg R^2 : 0.479**
  - **Distance week0 : 46.5 weeks**
  - **Avg MAE: 74.2**
- **Personal low confidence (n=1)**
  - **Avg R^2 : 0.394**
  - **Distance week0 : 13.0 weeks**
  - **Avg MAE: 128.3**

Problem → Is it Reliable?

We work with predicted values of Week 0.
We'll have a final prediction of the progression based on a starting prediction, there could be a big error propagation.

Possible other ways:

- **Work with predicted values at week 0 but weight data, giving much more importance to the 18 patients we know and less to the ones predicted (especially the ones with low confidence)**

- **Eliminate FVC at week 0 for everyone, use only CT scans as baseline (extracting from there the FVC?) therefore final output cannot be decline based on baseline FVC but a definite value**

- **Incorporate the prediction of the FVC baseline at time as an output of the model itself, so final output (FVC baseline, FVC 1year, FVC 2year, Progression Yes/No)**