

# Scientific Visualization Final Project

## SNF-Driven Patient Stratification and Vital Status Prediction

Nada Bou Kanaan

# Datasets

Clinical data Adenosquamous carcinoma patients: clinical data of 138 patients with Adenosquamous carcinoma with 24 features: age, sex, race, participant\_country, tumor\_site....additiol\_pathologic\_findings, bmi, alcohol\_consumption, tobacco\_smoking\_history, follow\_up\_days, 'vital\_status'..

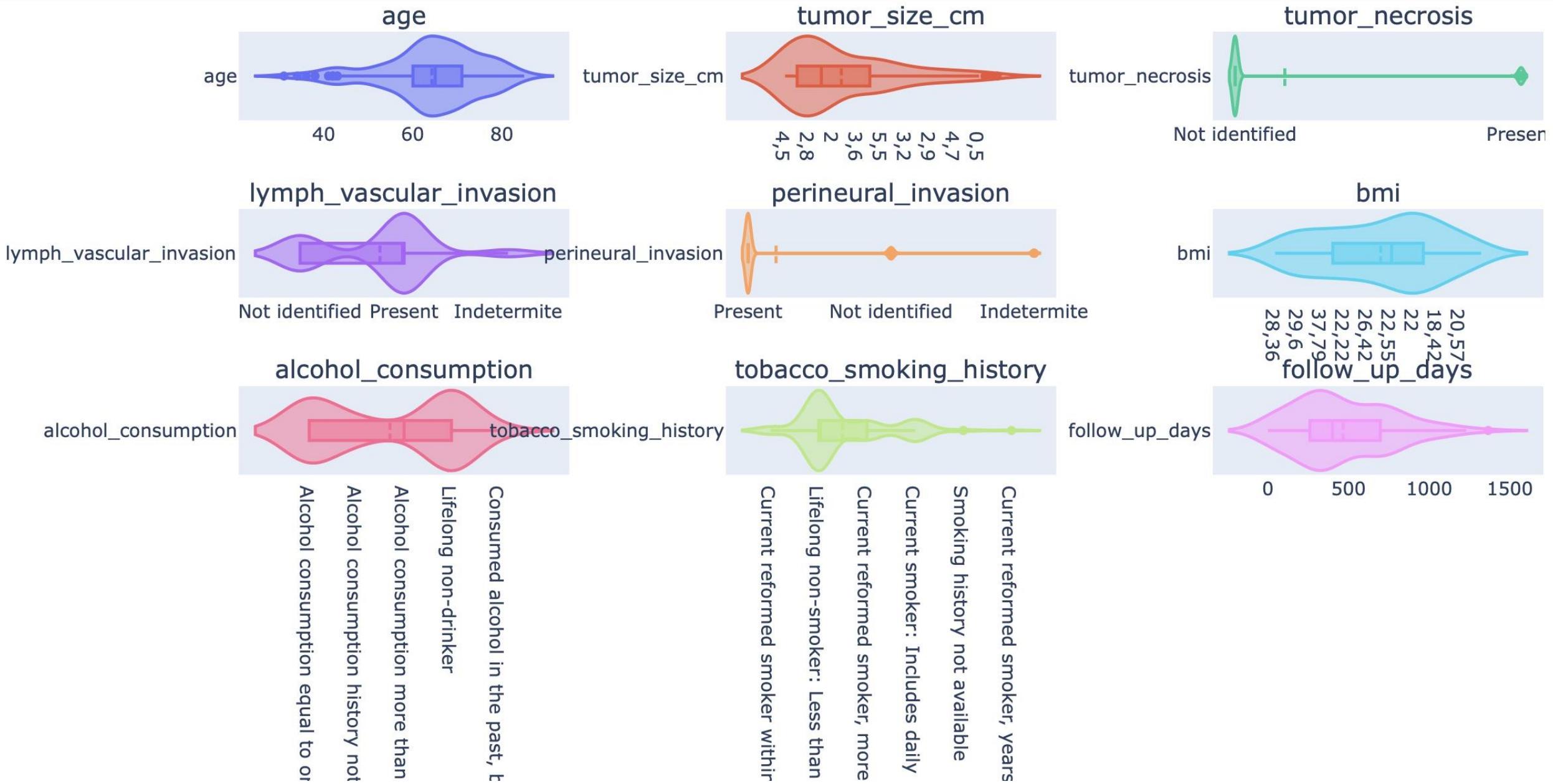
SCNA\_gene: gene level measurement of somatic copy number alterations for the same 138 patients

mRNA: mRNA expression data for the same 138 patients

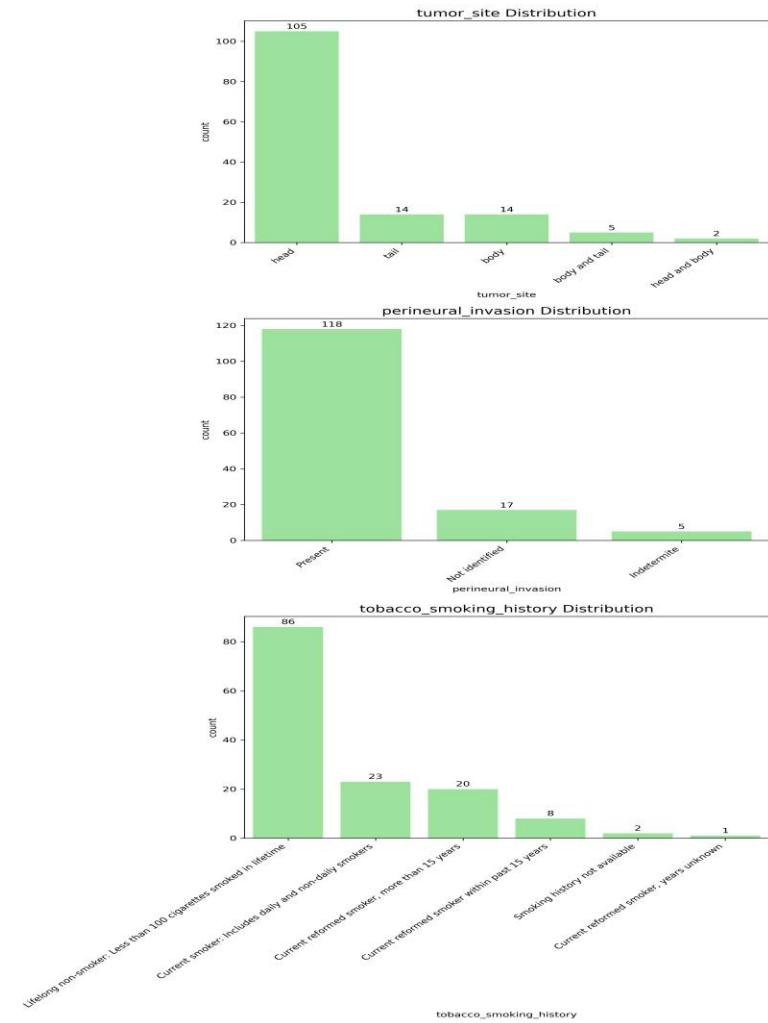
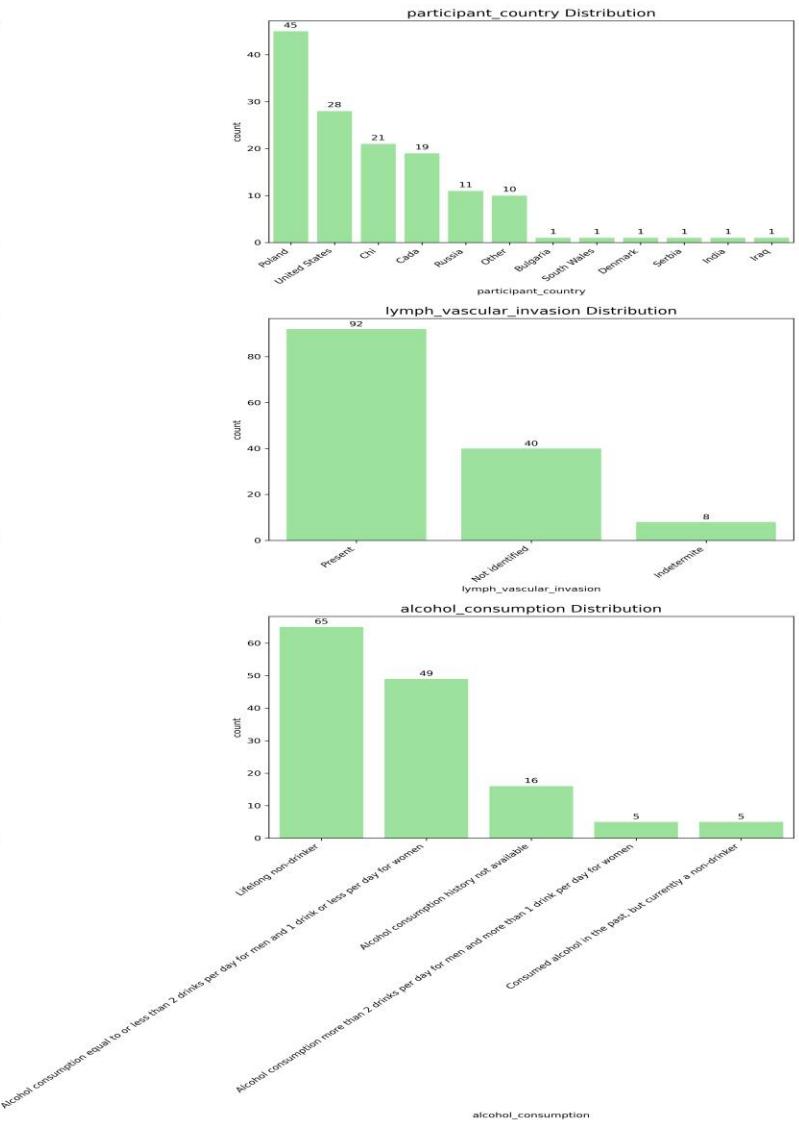
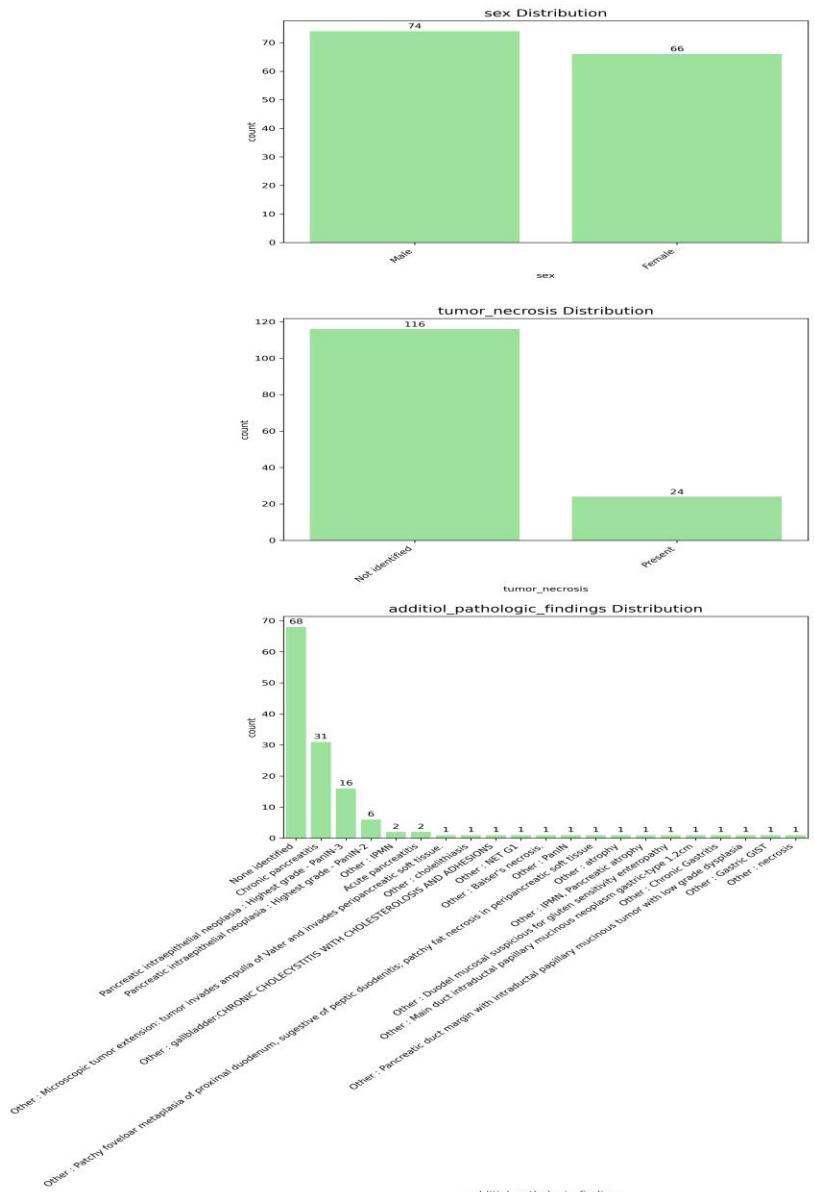
# First step: prediction without the genomics data

- This code analyzes the clinical data of 138 patients with Adenosquamous carcinoma. I trained two classifiers—Random Forest and K-Nearest Neighbors (KNN)—to predict the patients' vital status based on their clinical data. Finally, I used the models to predict the vital status for the four patients with missing information. I visualized the data composition and the feature importance to better understand the distribution of the clinical variables and which features most influenced the predictions.

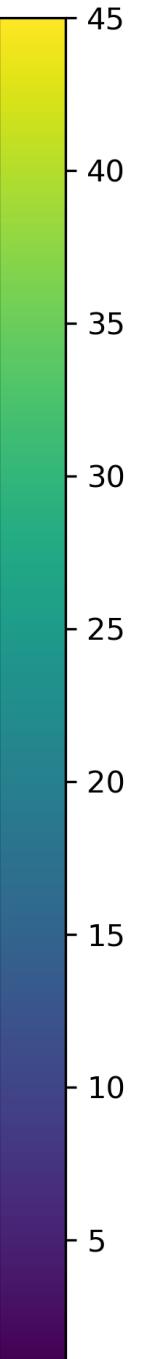
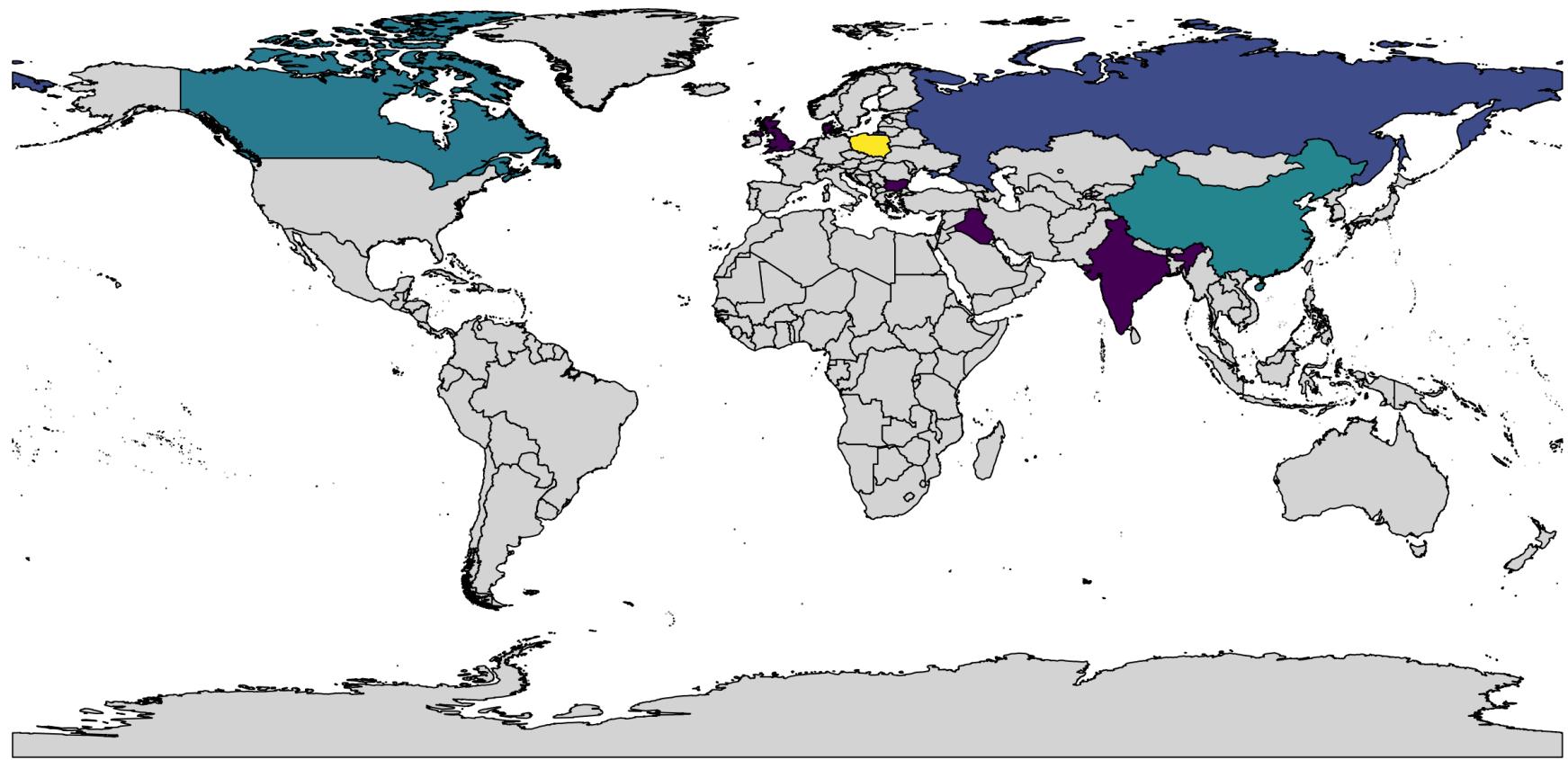
# Data analysing

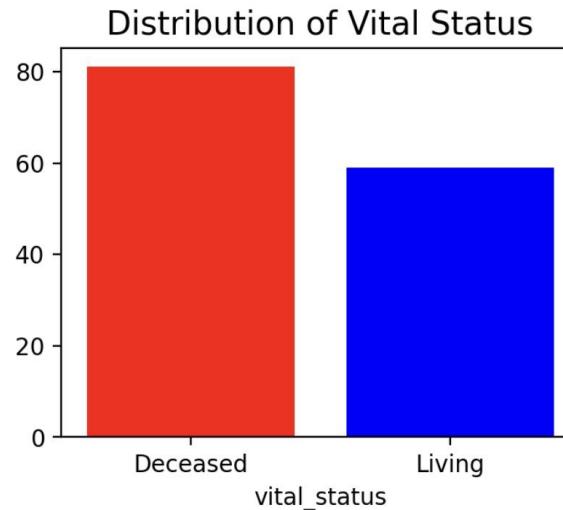
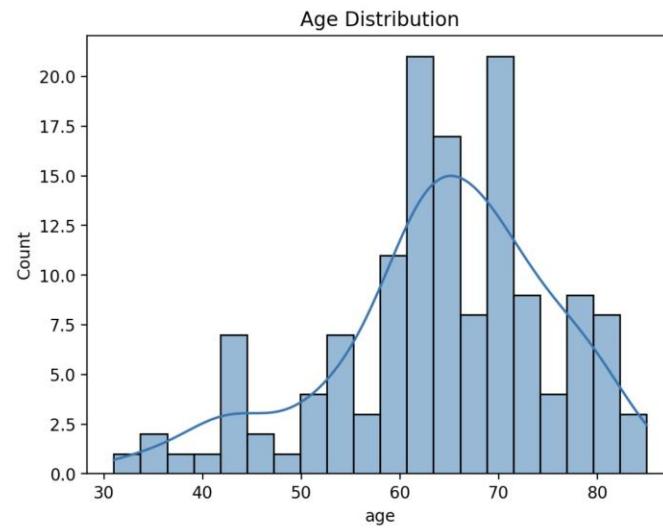
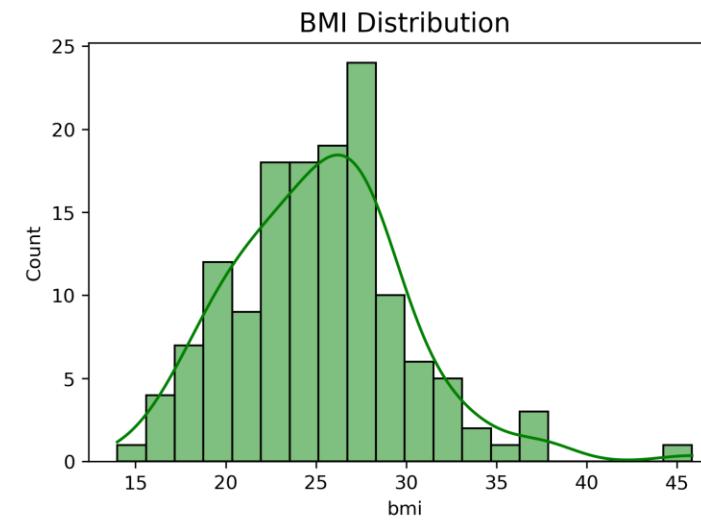
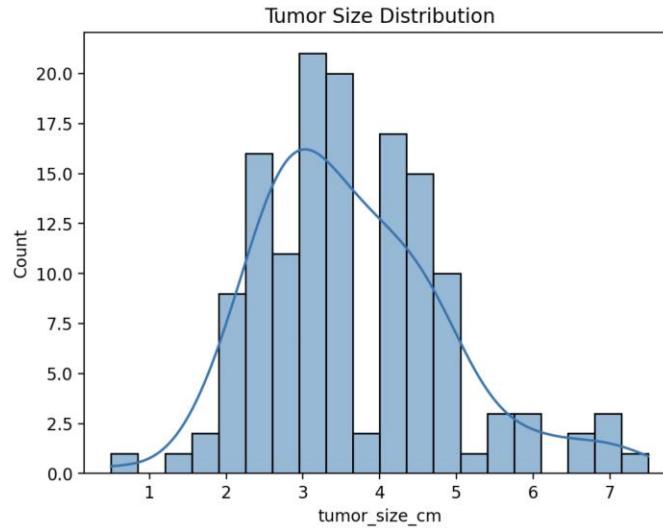


# Data analysing

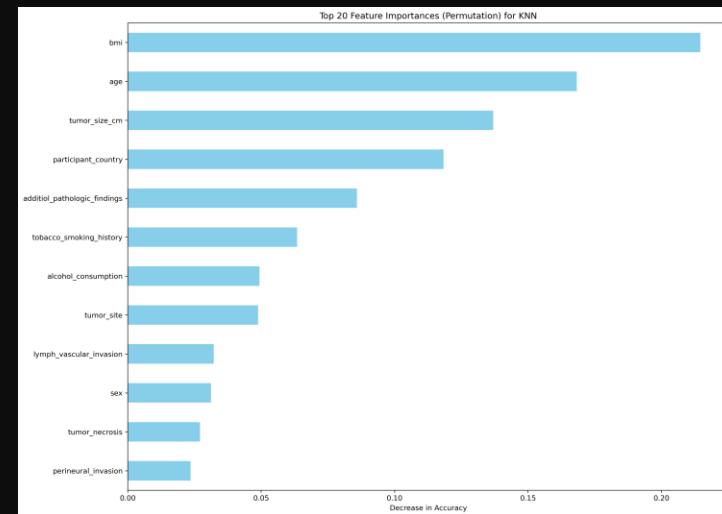
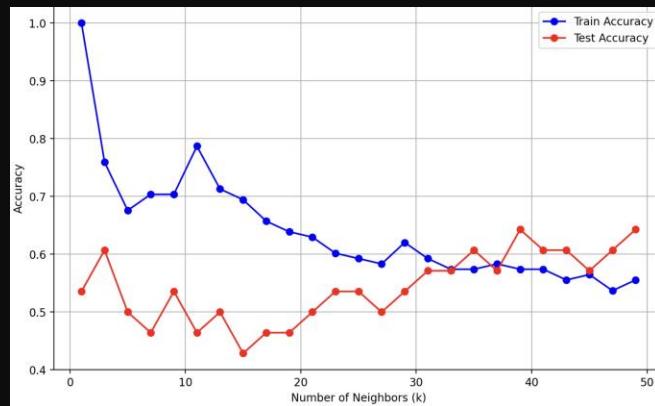


# Geographic Distribution of Clinical Samples





KNN feature  
importance: bmi, age  
optimal number of  
neighbors : best k 39



# Predictions

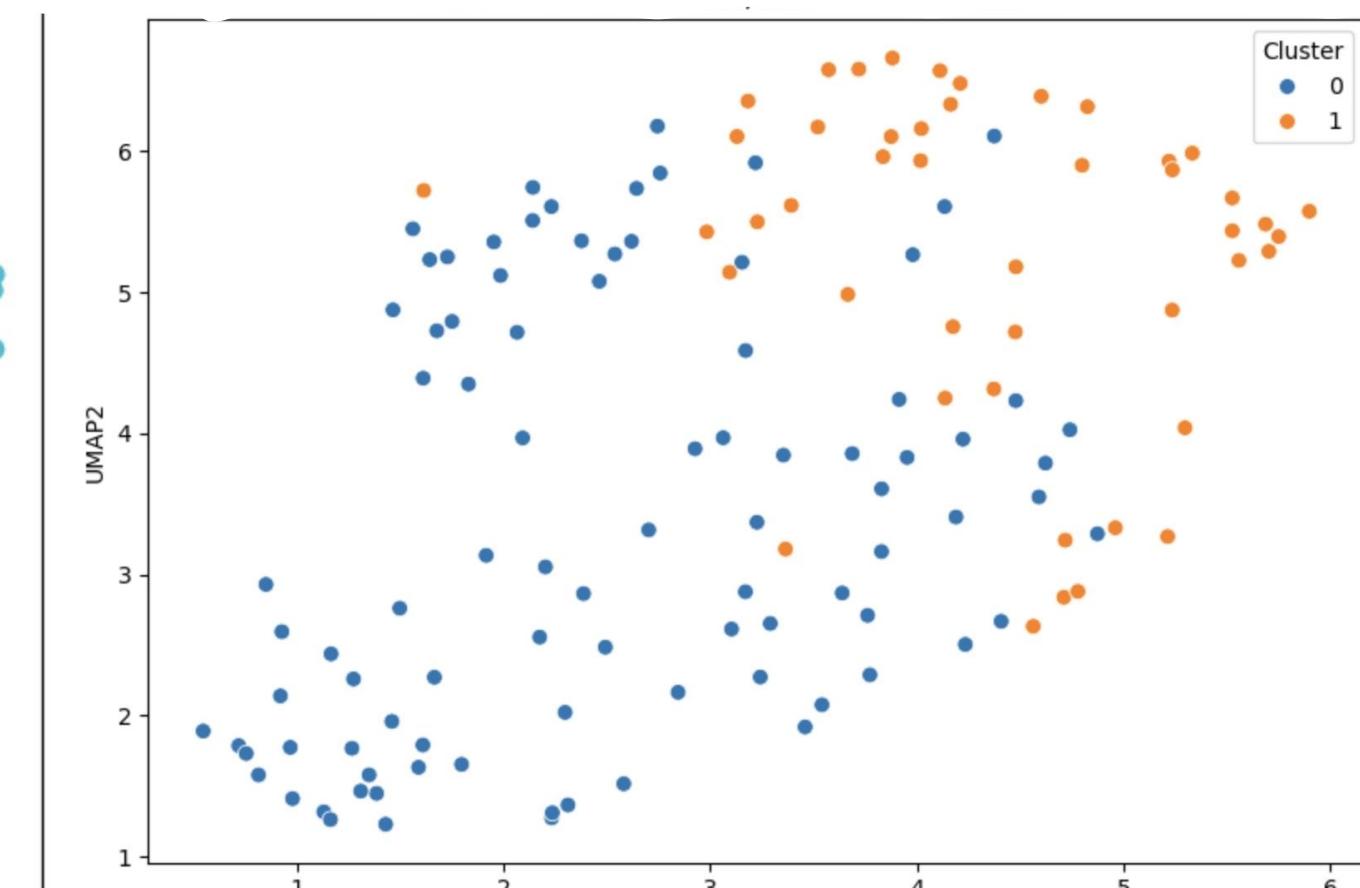
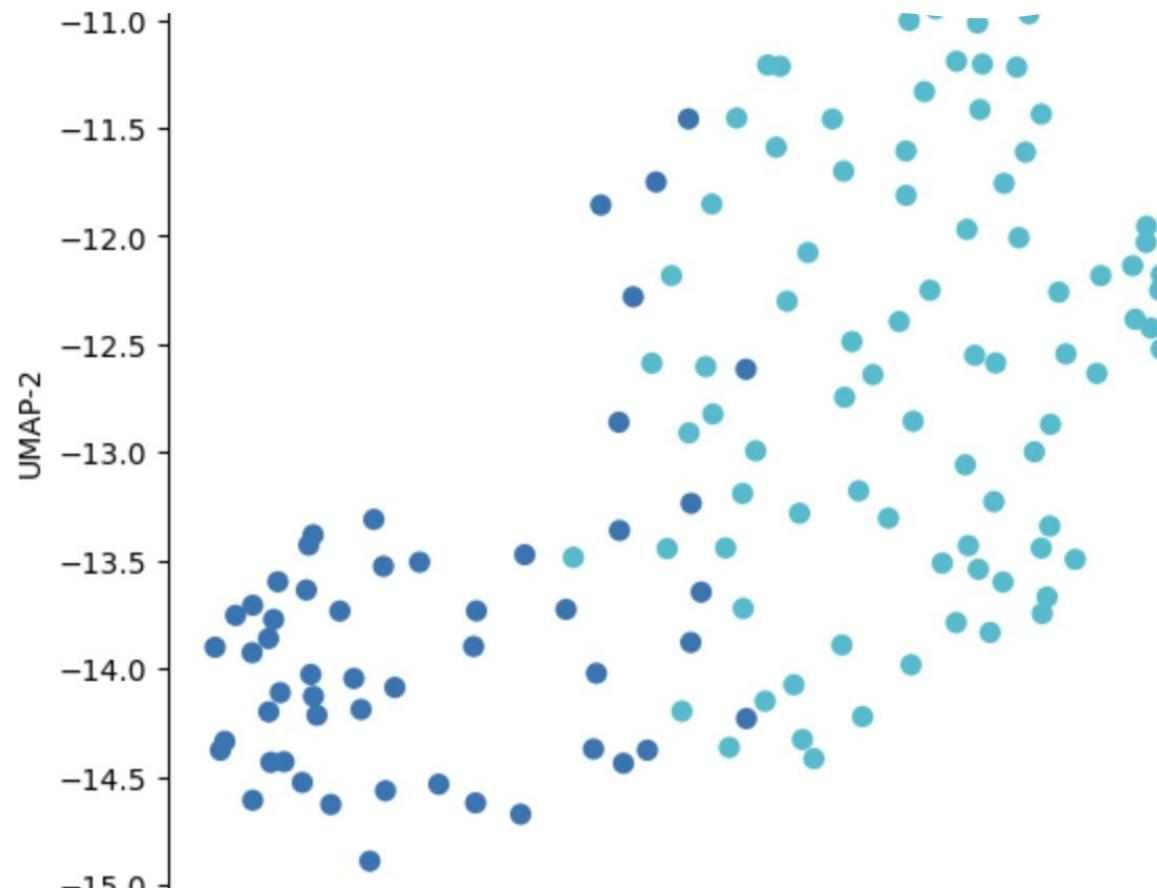
- case\_id vital\_status\_pred
- C3L-03395 1
- C3N-00709 0
- C3N-03039 0
- C3N-03069 0

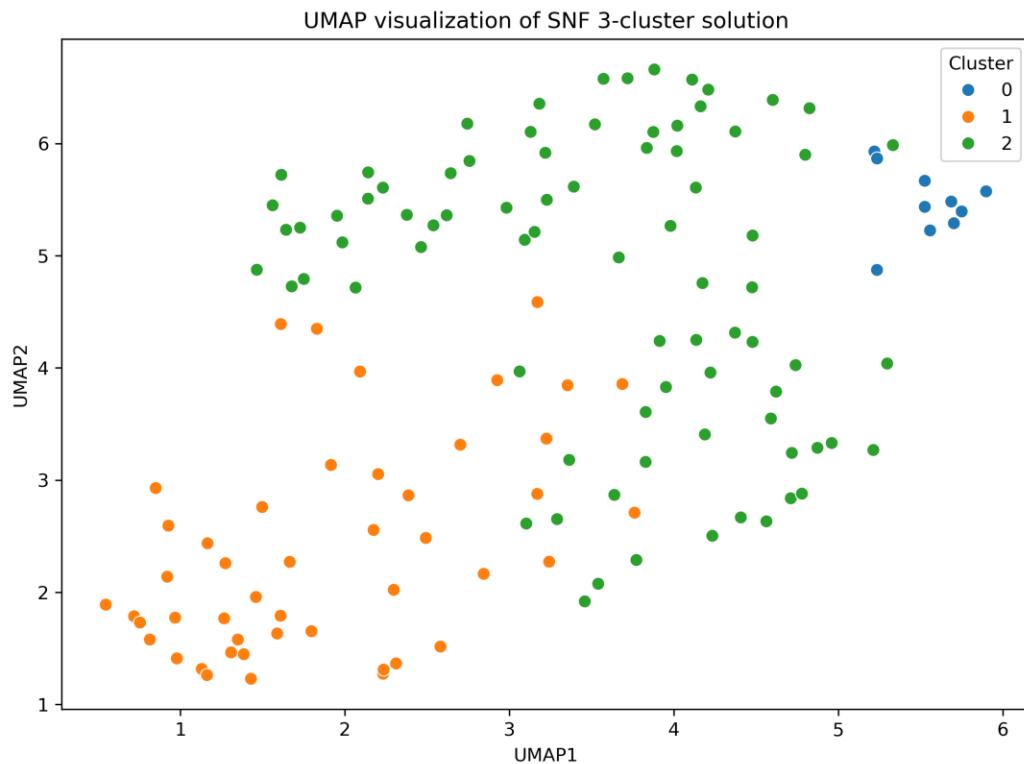
(0 is alive, 1 is deceased)

## Second step:

The SCNA\_gene and mRNA datasets were integrated using SNF in Python and subsequently clustered. The resulting cluster labels were added to the clinical data to perform the same classification, predicting the patients' vital status, and the results were compared with the previous classification outcomes.

- The SCNA\_gene dataset and the mRNA dataset were integrated using SNF in Python. After fusion, the data were transformed into a sample-by-sample similarity matrix and subsequently clustered. We initially used the parameters  $K = 20$  and  $\mu = 0.5$ , but this resulted in a low silhouette score. After optimizing the parameters, the clustering performance improved

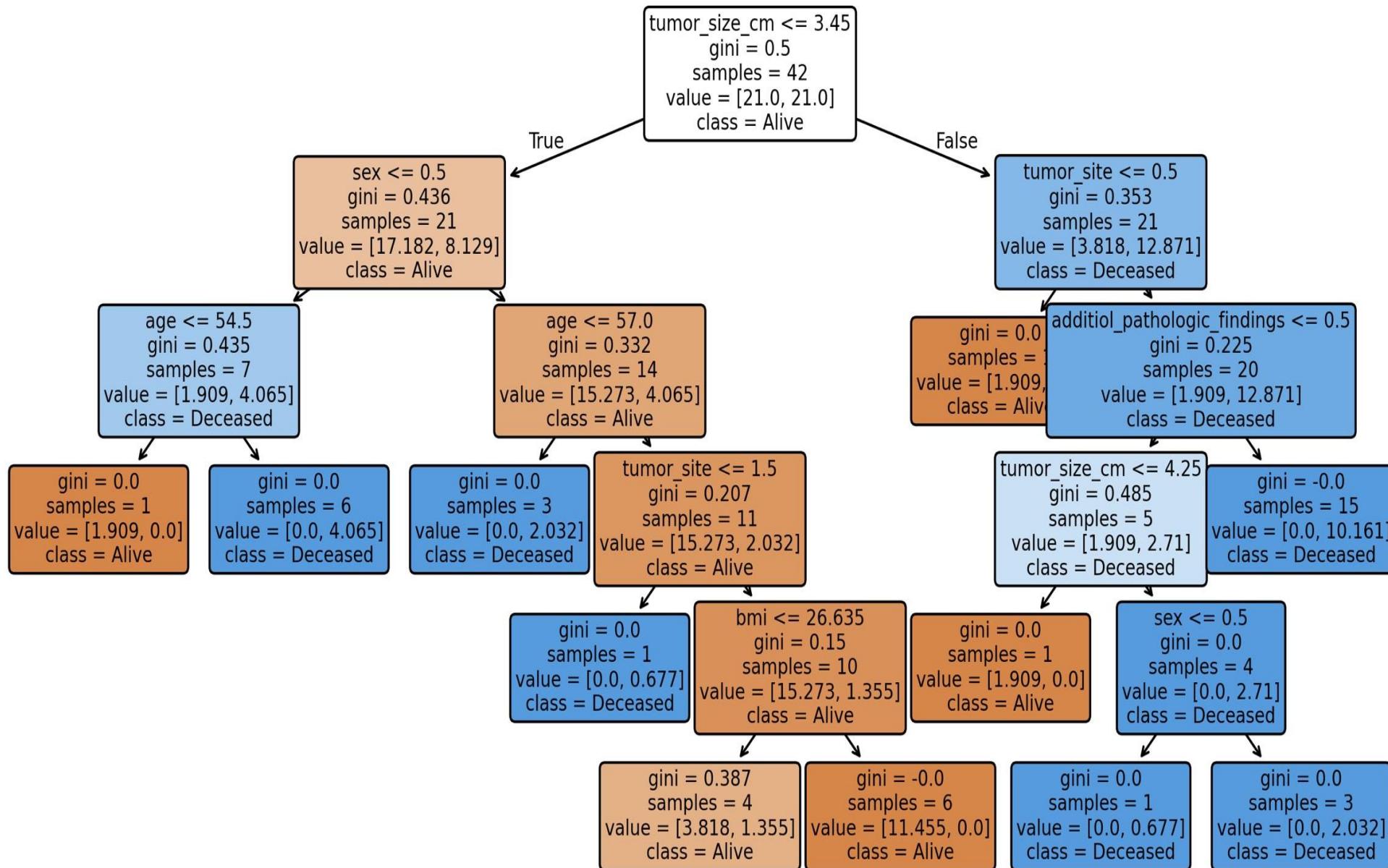




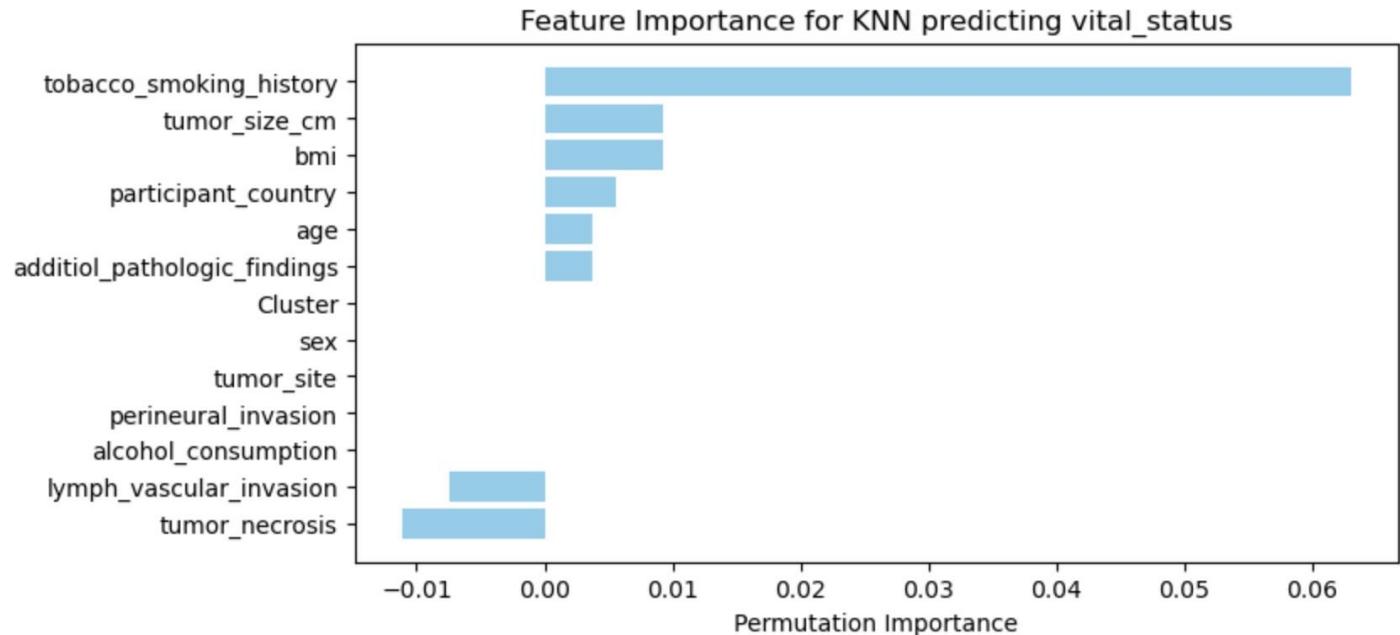
the number of clusters also affected the silhouette score, with the best score achieved when using three clusters.

clusters were still visually distinguishable even when the silhouette score was relatively low.

## Decision Tree for Predicting Vital Status



# KNN: integrated data



When the clusters were added to the classification, the results changed noticeably. The feature importance in the KNN model shifted substantially, with smoking and BMI remaining the dominant features

# Predictet vital status

case\_id predicted\_vital\_status predicted\_prob\_deceased

C3L-03395 1 0.8

C3N-00709 1 0.6

C3N-03039 1 1.0

C3N-03069 1 1.0

(0 is alive, 1 is deceased)