Breast-Cancer Attributes Prediction

Authors: | Yehonatan Ezra | Shay Morad | Lior Zats | Avi Kfir |

Part 3 – Unsupervised Learning

Goal: look for patterns in train.feats.csv that doctors might find interesting.

Principal-component analysis (PCA)

- first 9 components already explain ≈ 50 % of the total variance → engineered features do capture new information.
- biggest axis separates “large, surgically aggressive cases” from “small, early-stage cases”.

K-means clustering (k = 5) on the PCA space

- three big, partly overlapping clusters trace a smooth severity gradient.
- one tight cluster is dominated by triple-negative tumors (ER-, PR-, HER2-).
- another cluster groups most post-mastectomy visits.

t-SNE visualisation

- confirms that those five clusters are not random blobs: similar visits sit close, dissimilar ones are far.
- silhouette score ≈ 0. 1177 – modest but typical for high-dimensional clinical data.

Take-away: even with basic engineering and no labels, the data naturally organizes along biologically sensible dimensions (size, aggressiveness, subtype). These insights can guide future feature design and help clinicians spot outlier cases quickly.