ackage.
his data was made public in papers [1] and [2] that can be found at the references.

ll the data from the dataset was generated and measured from prostate tissues using urface-Enhanced Laser Desorption/Ionization (SELDI) Mass Spectrometry. It contains 654 ass spectra from 327 patients (each subject has 2 replicates). All patients are divided into groups; patients with prostate cancer (pca), patients with benign prostatic hyperplasia oph), control subjects (control). They took blood and serum samples from patients.
egarding the dimensionality of the data, it has 3 different classes so it is a multiclass ataset with 327 samples and 10523 features (points of spectra).

ethods

eplicate averaging and log transformation

a. Preprocessing steps
b. DOI: https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/log , base package
c. The preprocessing steps refers to everything you must do before ML, which is application dependent. Thus, in our case we have done replicate averaging (because we had two replicates per subject) and log transformation of the data to obtain a gaussian distribution.

ariance feature filter

a. Unsupervised feature filtering
b. Own programing. Attached R file submitted.
c. It is used to remove uninformative features from the analysis. The procedure consists in computing the variance of all features, sorting them and removing the features that explain a low variance of the data. In our case, we have optimized this threshold.

earest-Centroid Classifier

a. Classification algorithm
b. DOI: 10.1016/j.neuroimage.2005.06.017 , class package
c. The Nearest-Centroid Classifier is a classification algorithm used because of its simplicity and efficiency, especially when the classes are well separated. It is based on the following discriminant function, which will output a large value when the distance between the centroid and the feature vector is close (feature will correspond to the class that has the nearest centroid):

$$g_i(x) = -(x-\mu_i)^T(x-\mu_i)$$

-Nearest Neighbors

a. Classification algorithm
b. DOI: 10.1109/2.78568 , caret package

non-linearly separable. It operates on the principle that data points with similar features tend to belong to the same class. It consists in finding the “k” closest labeled examples in the training dataset and assigning the new feature vectors to the class that appears most frequently within the k-subset. Note that it is computationally and memory expensive, so we should not use it when we work with a huge dataset.

old-out validation

a. Data partition
b. DOI: 10.1111/j.2517-6161.1974.tb01479.x , caret package
c. The Hold-Out Method divides the dataset into two portions: a training set and a testing set. The principle is to train the model on the training set and evaluate its performance on the testing set. It allows for the assessment of how well the model generalizes to unseen data based on a single random split.

-Fold cross-validation

a. Validation technique
b. DOI: 10.1111/j.2517-6161.1974.tb01479.x , caret package
c. K-Fold Cross-Validation divides the dataset into k approximately equal-sized folds or subsets. The model is trained and tested k times, with each fold used as a testing set once. The performance is averaged over these iterations, reducing the impact of random variability. The principle is to assess the model's performance while ensuring that all data is used for both training and testing.

esults and Interpretation

ote that depending on the seed used to perform the analysis, we will obtain different esults. In fact, it could happen that we obtain a bigger CR in the external validation than in e internal validation. This could happen because there is an element of randomness ssociated with the data partition process. It could happen that the data in the external alidation is more similar to the training than the data in internal validation. Therefore, we get better balanced classification rate (BCR).

he initial step in our analysis involved the preprocessing of mass spectra data. By mplementing the replicate averaging, we reduced the impact of measurement variability ithin subjects. By implementing the log transformation we normalized the intensity stribution and highlighted major spectral peaks. This preparation of the data enhanced its uitability for subsequent machine learning analysis.

he workflow followed to implement a variance feature filter (unsupervised) consists in alculating the variance for each feature, sorting the obtained variances and visually ssigning a threshold to remove the features that show no significant variability in the full ataset. We have optimized this threshold in the internal validation by calculating the ariance for each feature, sorting the features, selecting the top 90% of features with higher ariance (we could also optimize the percentage of features that we keep, but we won’t do it) nd perform the analysis again.