Ovarian tumor subtype prediction using Gene Expression Data

Valentin Bernu
17/12

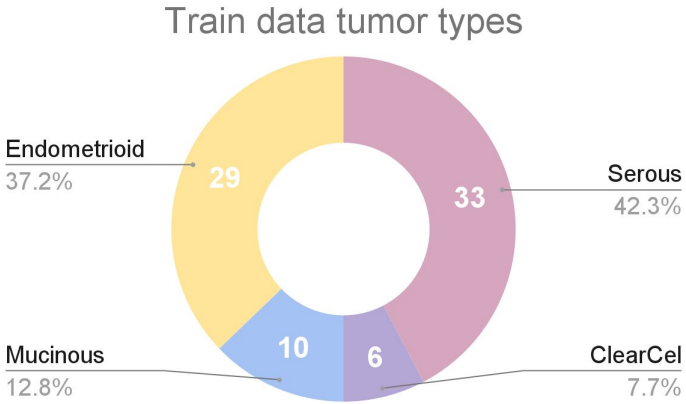
Plan

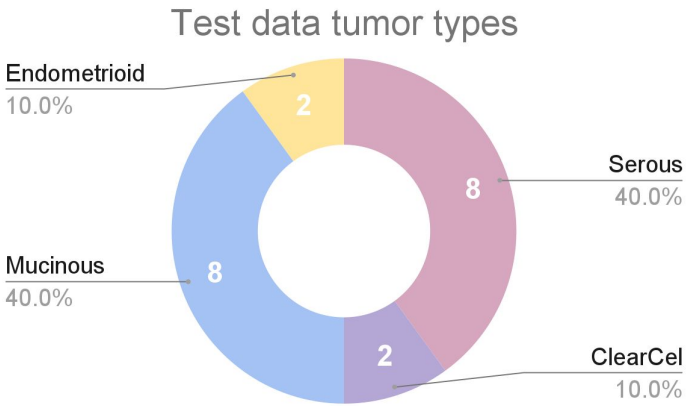
1. Data Exploration
2. Feature selection 1: Kruskal-Wallis test
3. Model benchmark
4. Feature selection 2: Sequential Feature Selection
5. Prediction & Evaluation

1. Data exploration

GSE6008 dataset

22286 columns **78** rows for training **20** rows for testing

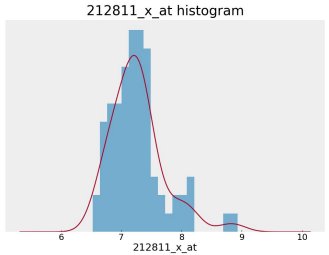


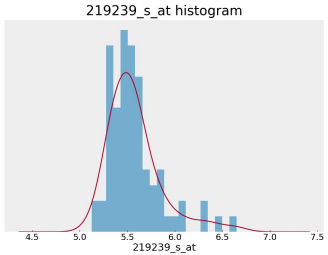


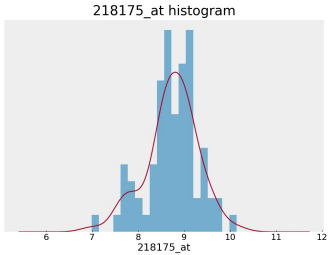
1. Data exploration

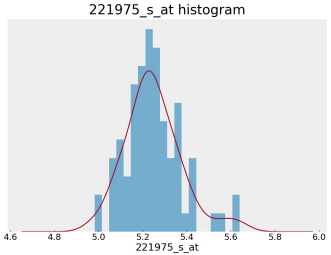
GSE6008 dataset

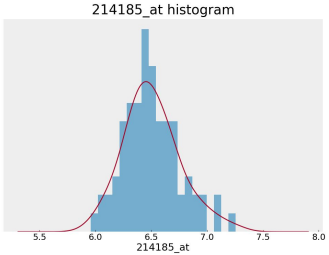
22283 gene expression columns

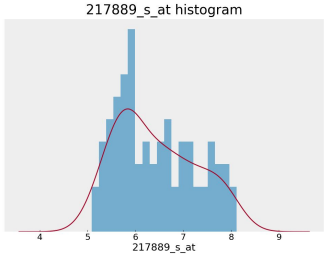






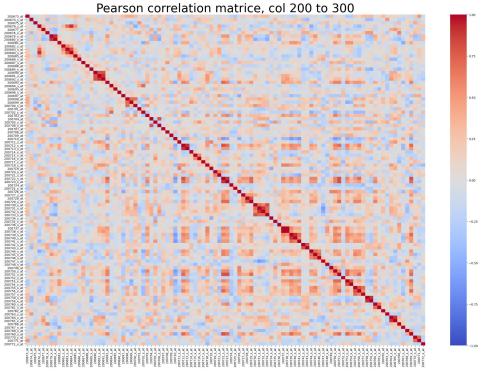


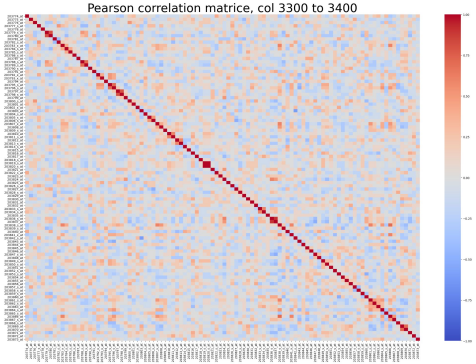


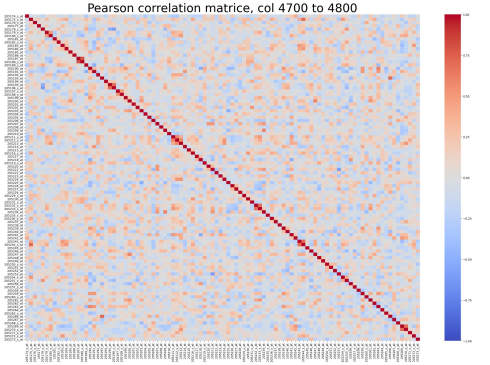


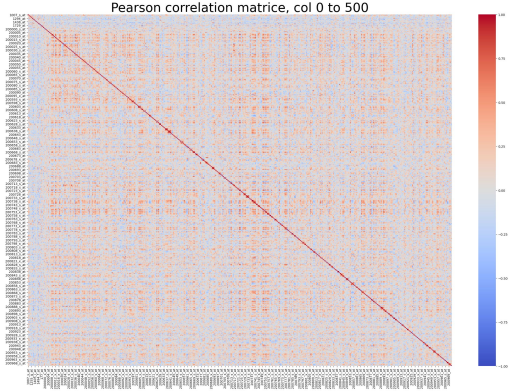
1. Data exploration

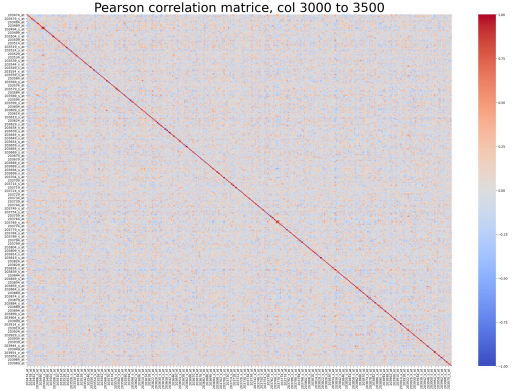
Correlation matrices

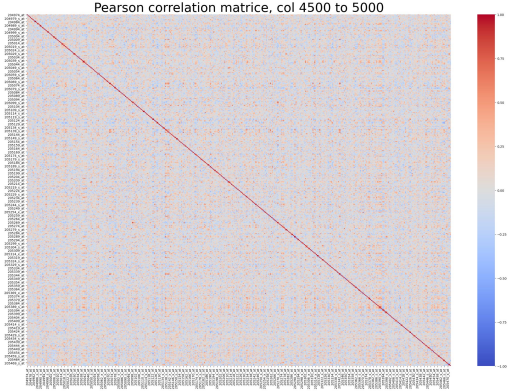








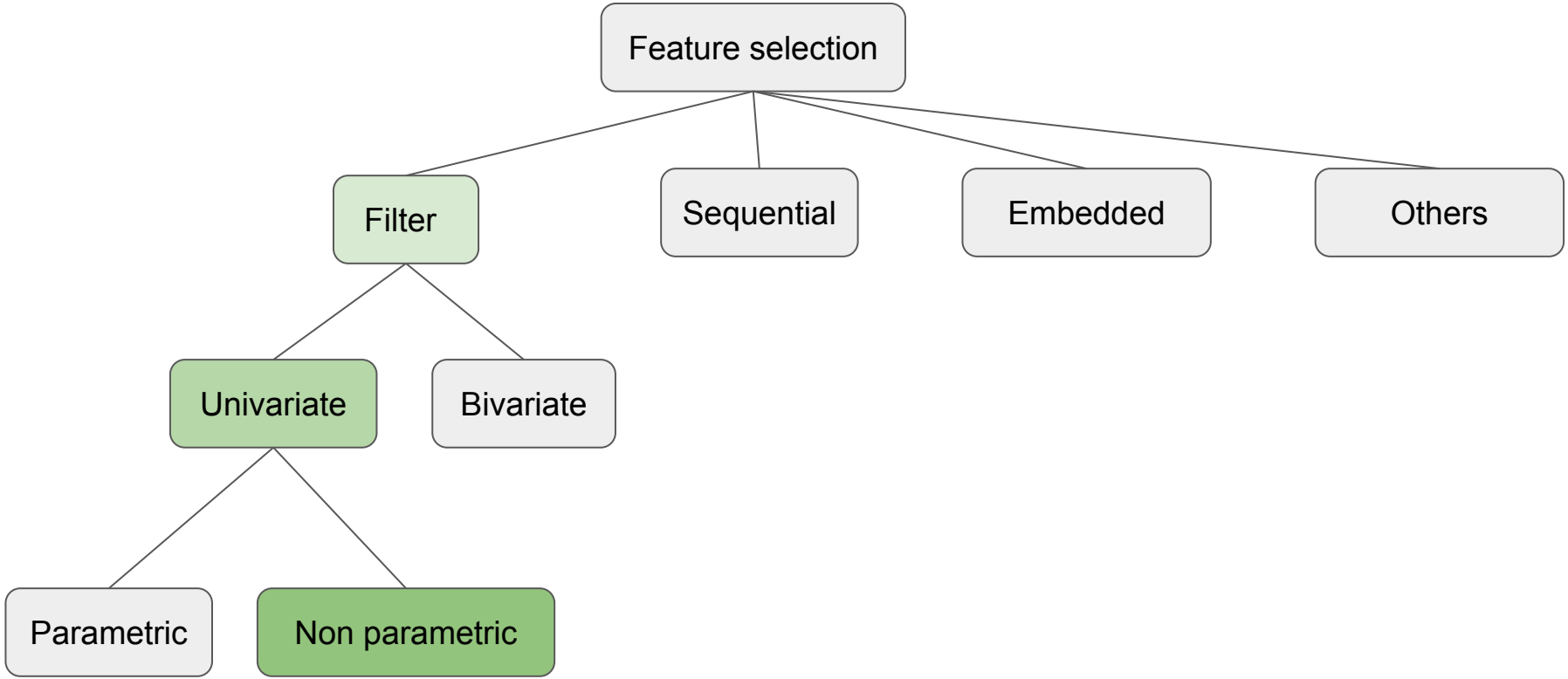




GSE6008 dataset

- **Tumor types differently distributed** in train and test data.
- Gene expression data present **different distribution shapes**, with **different means and standard deviations**.
- Normal distribution are frequent.
- Not strong correlation pattern detected in correlation matrices.

2. Feature Selection 1: Kruskal-Wallis test

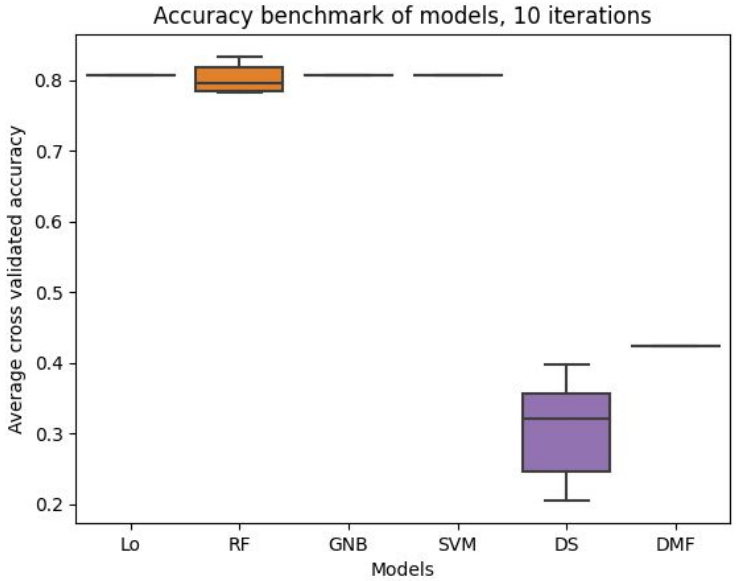


3. Model benchmark

- Logistic Regression (Lo)
- Random Forest (RF)
- Gaussian Naive Bayes (GNB)
- Support Vector Machines (SVM)
- Dummy Stratified (DS) *-generates predictions by respecting the training set's class distribution.*
- Dummy Most Frequent (DMF) *-always predicts the most frequent label in the training set.*

Method: 10 iterations with 3-fold cross validation

3. Model benchmark

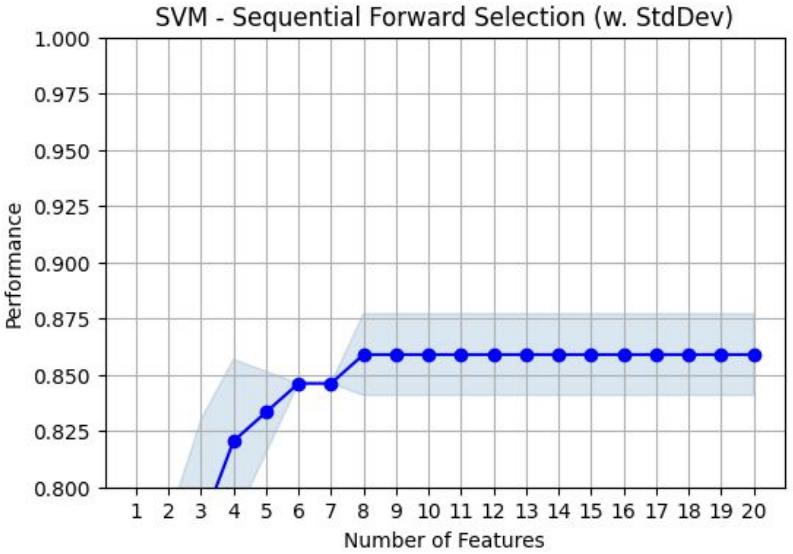


4. Feature Selection 2: Sequential Feature Selection (SFS)

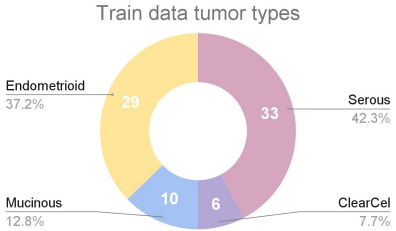
- What is SFS?

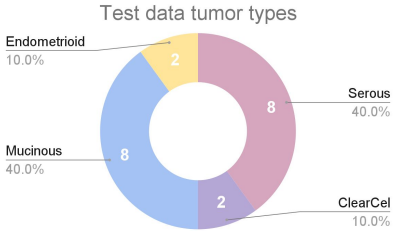
“At each stage, this estimator chooses the best feature to add or remove based on the cross-validation score of an estimator”, from scikit-learn documentation.

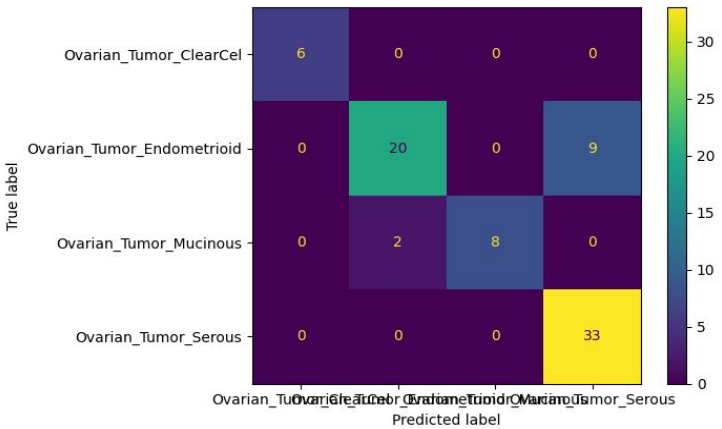
- Pros
  - Good performance
  - Robust to redundancy
- Cons
  - Requires high computational power
  - Can be trapped in local minima

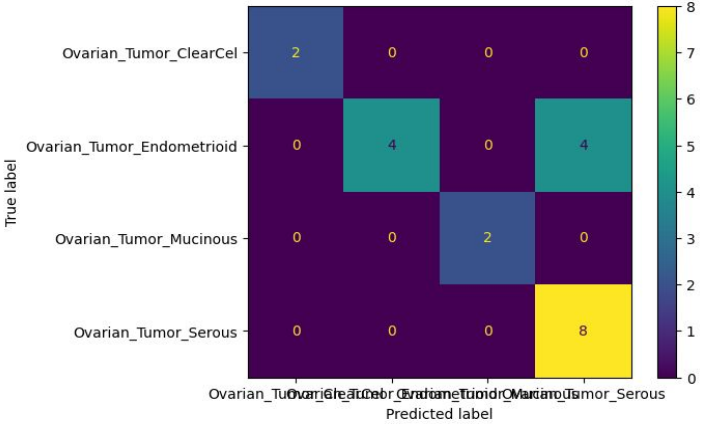


5. Performance evaluation









Future work directions

- Confidence score on predictions
- Study of selected features
- Class imbalance handling
- Regularization (embedded feature selection)

Thank you!