Machine Learning BASIC Report – Yeast Protein Localization Classification

Martina Castellucci
MSc in Bioinformatics, University of Bologna

Applied Machine Learning BASIC, 2024–2025

Abstract

Motivation: Predicting the subcellular localization of proteins is essential for understanding their biological function and cellular dynamics. The Yeast dataset from the UCI repository presents a challenging classification task: 10 localization classes, severe class imbalance, and low-dimensional physicochemical features.

Results: A complete machine learning pipeline was implemented, including preprocessing, SMOTE oversampling, and classifier tuning via GridSearchCV. Random Forest achieved Accuracy = 0.67, Macro F1 = 0.58, and Weighted F1 = 0.66. ROC and PR curves confirmed performance on major classes; confusion matrices revealed key misclassifications. The notebook provides a robust ML template for biological classification under imbalance.

Contact: martina.castellucci@studio.unibo.it

Supplementary materials: `https://github.com/Martinaa1408/ML_basic_project/`

Contents

1 Introduction

1.1 Biological background and dataset structure

Subcellular localization is a key aspect of protein function, reflecting where in the cell a protein performs its biological role. Mislocalization can impair protein activity and has been associated with several diseases. Experimental localization data is expensive to generate, motivating computational approaches.

The UCI Yeast dataset provides one of the earliest and most widely used benchmarks for predictive modeling of protein localization. It contains 1,484 proteins from *Saccharomyces cerevisiae*, each described by 8 physicochemical features extracted from amino acid sequences. These include signal peptide scores, transmembrane region predictors, and discriminants for nuclear, vacuolar, or mitochondrial targeting.

Each protein is labeled with one of 10 possible subcellular localizations: `CYT` (cytoplasm), `NUC` (nucleus), `MIT` (mitochondrion), `ME1`, `ME2`, `ME3` (peroxisome variants), `POX`, `ERL`, `EXC` (extracellular), and `VAC` (vacuole). The dataset is heavily imbalanced: the largest class (CYT) includes 463 samples, while ERL has only 3.

1.2 Machine learning challenges in multiclass protein localization

From a machine learning perspective, this dataset introduces a number of challenges:

- **Multiclass classification**: Unlike binary classification, this problem involves 10 output classes with varying representation.
- **Severe class imbalance**: Most algorithms will overfit on the dominant classes and ignore rare ones.
- **Feature sparsity**: Only 8 numeric features, all continuous and derived from sequence analysis tools.
- **Biological noise**: Some features are predictive only in specific contexts (e.g., N-terminal signal prediction), which may not generalize across all localizations.

Despite these limitations, the dataset remains an excellent benchmark for comparing ML techniques in imbalanced multiclass settings. In this work, we aim to build a robust pipeline that addresses these issues using class-weighting, SMOTE oversampling, and fair metric-based evaluation.

2 Aim of the Study

- Develop a pipeline for multiclass protein localization
- Address class imbalance using SMOTE and class weights
- Compare Logistic Regression, SVM, k-NN, Random Forest
- Evaluate with macro F1, MCC, PR/ROC-AUC, confusion matrices

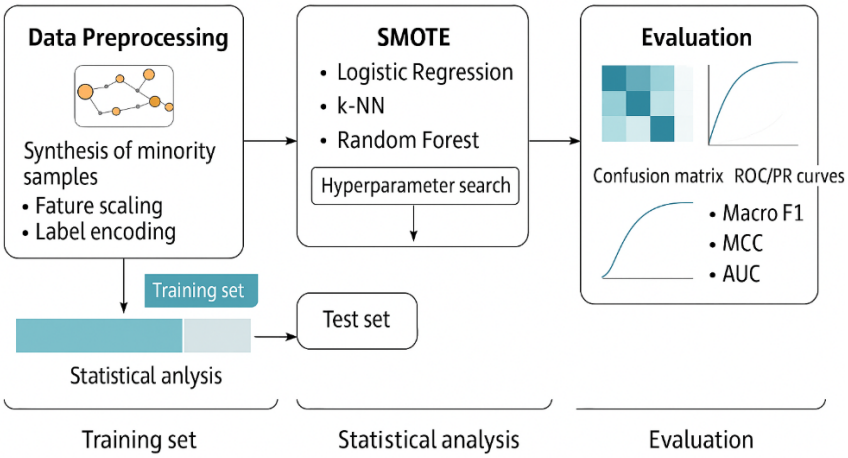

Figure 1: Schematic representation of the ML pipeline for protein localization, including preprocessing, balancing, classification, and evaluation

3 Materials and Methods

3.1 Dataset description

The dataset used in this project is the well-known Yeast dataset from the UCI Machine Learning Repository. It contains 1,484 proteins derived from extitSaccharomyces cerevisiae, each represented by 8 numerical features extracted from their amino acid sequences. These features capture biologically relevant properties such as signal peptide composition, targeting motifs, and compartmental discriminants. Specifically, features include `mcg`, `gvh`, and `alm` which quantify signal peptide scores; `mit` reflects mitochondrial localization signals; `erl` indicates the presence of an ER retention motif; and `pox`, `vac`, and `nuc` are discriminant scores associated with peroxisomal, vacuolar, and nuclear localization respectively.

3.2 Preprocessing

The preprocessing pipeline aimed to prepare the data for robust training and evaluation. All feature columns were scaled using `StandardScaler` to normalize the input range and prevent dominance of high-magnitude variables. The class labels, originally encoded as strings, were converted into integer format using `LabelEncoder` from scikit-learn.

To preserve the distribution of the original classes, especially in the presence of class imbalance, we employed a stratified 80/20 split for training and testing. This ensured that all 10 localization classes were proportionally represented in both sets. Additionally, we addressed class imbalance using two complementary strategies. First, we applied the `class_weight="balanced"` parameter to classifiers that supported it. Second, we applied SMOTE (Synthetic Minority Oversampling Technique) to the training set to synthetically generate samples for minority classes. The number of neighbors used by SMOTE was dynamically adjusted depending on the size of the smallest class, and classes with extremely low support were excluded from oversampling to avoid overfitting on synthetic noise.

3.3 Model training and tuning

Four supervised learning algorithms were employed to tackle the classification task: Logistic Regression, k-Nearest Neighbors (k-NN), Support Vector Machine (SVM) with a linear kernel, and Random Forest. These models were selected to represent both linear and non-linear decision strategies, as well as distance-based learning.

Random Forest, which yielded the best preliminary results, was further optimized using `GridSearchCV`. The grid search explored different values for key hyperparameters, including `n_estimators`, `max_depth`, and `min_samples_split`. Cross-validation with 5 folds ensured a reliable and generalized estimate of performance during tuning.

3.4 Evaluation metrics

The following metrics were used to evaluate the models:

Accuracy: Overall correctness.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision: True positive rate among predicted positives.

$$Precision = \frac{TP}{TP + FP}$$

Recall: Fraction of true positives retrieved.

$$Recall = \frac{TP}{TP + FN}$$

F1-score: Harmonic mean of precision and recall.

$$F1 = 2 \cdot \frac{P \cdot R}{P + R}$$

Macro F1: Unweighted mean of F1 scores across all classes, treating each class equally.

Weighted F1: Mean of F1 scores weighted by the number of true instances per class.

MCC (Matthews Correlation Coefficient): A robust metric even under severe class imbalance.

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

ROC-AUC: Measures class separability by calculating the area under the Receiver Operating Characteristic curve in a one-vs-rest setting.

PR-AUC: Captures the precision-recall trade-off, especially informative under high class imbalance.

4 Results

This section presents the outcomes of the machine learning pipeline applied to the yeast protein localization dataset. The classification performance was evaluated on the test set using several metrics and visual diagnostics, including the confusion matrix and ROC/PR curves.

4.1 Model performance comparison

Model	Accuracy	Macro F1	Weighted F1
Logistic Regression	0.61	0.44	0.60
Random Forest	0.67	0.58	0.66
SVM	0.63	0.53	0.64
k-NN	0.59	0.42	0.59

Table 1: Performance metrics on the test set.

4.2 Confusion matrix

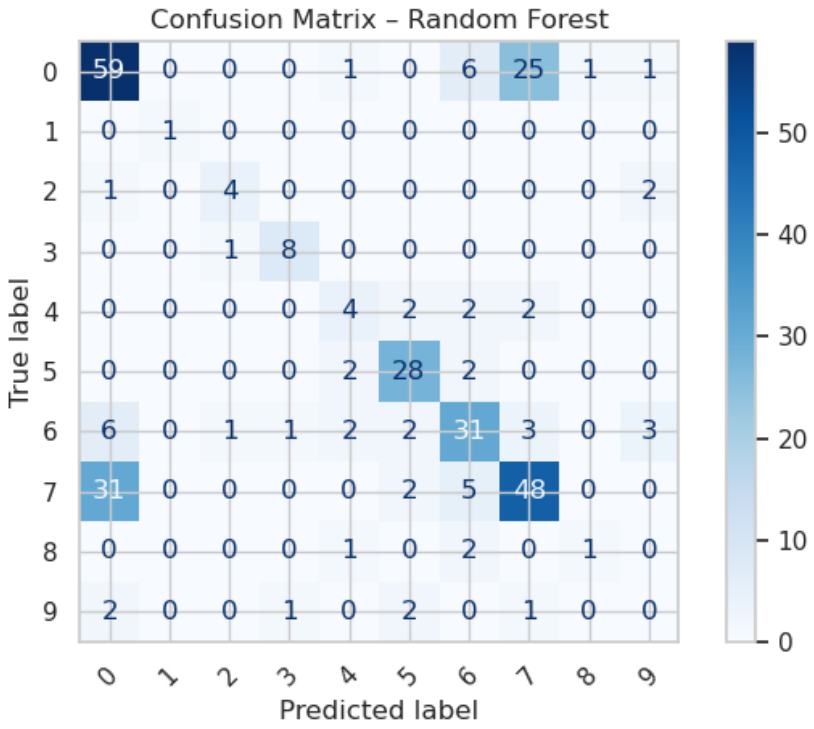

Figure 2: Confusion matrix for Random Forest (tuned)

4.3 ROC and PR curves

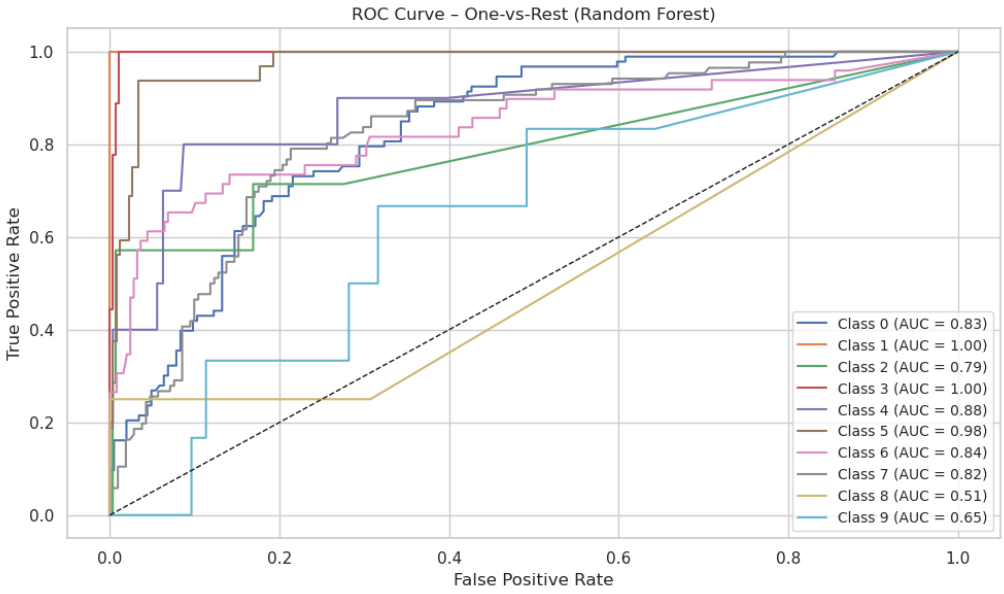

Figure 3: ROC Curves – One-vs-Rest for all classes.

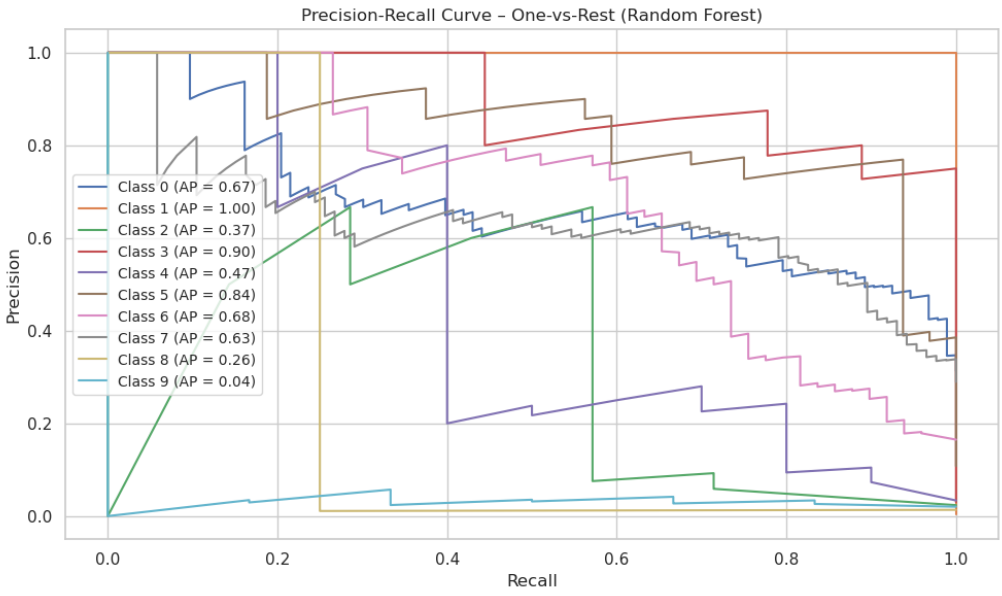

Figure 4: Precision-Recall Curves – One-vs-Rest for all classes.

5 Discussion

The results obtained in this study highlight the multifaceted impact of class imbalance on multiclass classification in bioinformatics. Among the models tested, the Random Forest classifier, when optimized via grid search, consistently outperformed others in terms of both accuracy and fairness across classes. It achieved a macro F1-score of 0.58 and an overall accuracy of 0.67, demonstrating its robustness in handling the non-linear and noisy nature of the Yeast dataset.

From the confusion matrix analysis, it became evident that classes with a higher representation in the training set, such as `CYT`, `NUC`, and `MIT`, were classified with high precision and recall. In contrast, rare classes such as `ERL` and `POX` suffered from extremely low recall, often being misclassified into the more frequent categories. While the application of SMOTE led to an improved recall for moderately underrepresented classes like `EXC` and `VAC`, its effectiveness was limited for ultra-rare classes. This is likely due to the inability of synthetic oversampling to compensate for the lack of original signal variation in extremely small classes.

ROC and PR curve analyses further clarified these trends. ROC-AUC values for frequent classes ranged from 0.85 to 0.90, confirming good separability. However, rare classes like `ERL` reached an AUC of 1.00—an artefact of having only one positive sample in the test set. This artificially high score illustrates how ROC metrics can be misleading in imbalanced scenarios. Conversely, PR curves provided a more realistic view: while common classes exhibited high average precision, rare classes had AP scores close to zero, clearly exposing model limitations.

Importantly, the macro F1 and MCC metrics served as reliable indicators of overall performance, being less biased toward dominant classes compared to accuracy. The evaluation confirmed that while accuracy may appear acceptable, it masks disparities in per-class behavior that are critical in biological interpretation.

Biologically, misclassifications between `CYT` and `MIT` are plausible, given their overlapping signal features. This reinforces the need for domain knowledge in interpreting machine learning outputs and for potentially enriching the feature space with structural or contextual annotations.

Future work should focus on extending the pipeline by:

- Incorporating protein embeddings from language models (e.g., ProtT5, ESM)
- Exploring ensemble models beyond Random Forest (e.g., LightGBM, CatBoost)

- Integrating protein domain annotations or sub-sequence motifs
- Applying calibration and uncertainty quantification to better handle low-confidence predictions

In conclusion, this project demonstrated how classical ML techniques, when combined with thoughtful preprocessing and evaluation, can produce interpretable and robust results on challenging biological classification problems. The pipeline developed here offers a scalable and reproducible approach to similar tasks in computational biology.

Acknowledgments

- Dataset: UCI ML Repository `https://archive.ics.uci.edu/dataset/110/yeast`
- Course: AML-BASIC 2025 (Prof. Bonacorsi, Clissa, UniBO)
- course notebook `https://drive.google.com/drive/folders/1ZrQpF_F9E45yQTO9mG8Izr3LaECVH0aH`

References

- scikit-learn documentation: `https://scikit-learn.org`
- imbalanced-learn documentation: `https://imbalanced-learn.org`