BRAC UNIVERSITY
Faculty of Computer Science and Engineering

CSE713– TASK #7
ADVANCED SYNTACTIC PATTERN RECOGNITION

Instructor: Annajiat Alim Rasel

2024/09/28

Name: Shaoun Chandra Shill
Student ID: 23373005

In machine learning with scikit-learn, various pitfalls can degrade model performance or misrepresent results if not handled carefully. Common issues include inconsistent preprocessing and data leakage. Preprocessing steps like scaling or feature extraction should be applied consistently to both training and test datasets; failing to do so can lead to suboptimal performance. Pipelines, which chain preprocessing and modeling steps, help mitigate this by automating these processes.

Data leakage, another frequent issue, happens when information from the test data influences model training. This leads to overly optimistic performance during cross-validation, but poor generalization on unseen data. To prevent this, preprocessing, such as feature selection, must only consider training data, not the test set.

Another important aspect of scikit-learn involves handling randomness. Estimators and cross-validation splitters use randomness in various stages, and controlling this randomness via the random state parameter ensures reproducibility. Proper usage, like passing integer seeds for consistent results or leaving it flexible for robustness, is crucial for reliable outcomes across different executions and splits.

By adopting pipelines and carefully controlling random processes, users can avoid common pitfalls and improve the quality of their machine learning models.