

We can see some challenges mentioned here while working with scikit-learn.

A sig Uniform pre-processing is crucial for improving model performance. Any adjustments made to train data, such as scaling, must be replicated on test data to avoid negative results. Pipelines improve automation and prevent data leakage, which can lead to heavily optimistic outcomes when training models using test data. Feature selection should be based on training data, not test data. Managing uncertainty is another challenge. Using an integer to specify 'random state' allows for repeatable outcomes but may cause instability in cross-validation. Using these instances adds variety and enhances the model's tolerance to anomaly through testing in various scenarios. Some estimators, such as 'Random forest' or 'Cross-validation Splitter' respond differently based on their 'random state', resulting in inconsistent results between folds due



marico

DATE:

S	M	T	W	T	F	S
---	---	---	---	---	---	---

to their variability. Avoid using 'random_state = None' and follow recommended practices for dealing with unpredictability in CV processes, including the distribution of Random Forest Classifier' or other random estimators.