

One concern here is that if a learning-based method like PCA is used for the dimensionality reduction on the dataset. If PCA is trained on the training set, then any cross-validation split that uses some portion of the training dataset will cause data leakage?