Breakdown of sklearn.datasets.make_classification

This function generates a random classification problem.

1. **n_samples** (int, default=100):
    - Sample Size
    - Description: The number of samples to generate.
    - Example: If I want to simulate a study with 1000 individuals, set `n_samples=1000`.

2. **n_features** (int, default=100):
    - Number of Predictor Variables
    - Description: The number of features (independent variables) in the dataset.
    - Example: If I was simulating a situation with 5 different predictors (like gender, major, etc.), I would set `n_features=5`.

3. **n_informative** (int, default=10):
    - Number of Informative Features
    - Description: The number of features actually used to build the model.
    - Example: If only 3 out of 5 features are relevant to my prediction, I would set `n_informative=3`.

4. **n_redundant** (int, default=2):
    - Number of Redundant Predictors
    - Description: These features are linear combinations of the informative features and do not provide additional information for classification.
    - Example: To have 3 features that are combinations of the informative ones, set `n_redundant=3`.

5. **n_repeated** (int, default=0):
    - Number of Duplicated Variables
    - Description: The number of duplicated features, drawn randomly from the informative and the redundant features; introduces no new information.
    - Example: For 2 duplicated features, use `n_repeated=2`.

6. **n_classes** (int, default=2):
    - Number of Categories/Classes
    - Description: The number of classes (or labels) in the dataset.
    - Example: For a multi-class problem with 4 classes, I would set `n_classes=4`.

7. **n_clusters_per_class** (int, default=2):
    - Description: The number of clusters per class.
    - Example: If each class is composed of 3 clusters, use `n_clusters_per_class=3`.

8. **class_sep** (float, default=1.0):
    - Class separation
    - Description: How far apart the classes are on a plane. Larger values spread out the clusters/classes and make the classification task easier.
    - Example: For more separated classes, I would increase `class_sep` above 1.0.

9. **flip_y** (float, default=0.01):
    - Label noise
    - Description: The fraction of samples whose class is flipped randomly. It adds noise and makes the classification task harder.
    - Example: For a noise level of 5%, I would set `flip_y=0.05`.

10. **weights** (list of floats or None, default=None):
    - Description: The proportions of samples assigned to each class. If None, the classes are balanced.
    - Example: For a dataset with 70% in one class and 30% in another, I would use `weights=[0.7, 0.3]`.

11. **random_state** (int, RandomState instance or None, default=None):
    - The random seed
    - Example: For reproducible results, set `random_state` to a fixed number like `random_state=123`.

12. **shuffle** (bool, default=True):
    - Randomization
    - Description: Shuffle the samples and the features.

13. **scale** (float or array of floats, default=1.0):
    - Description: Scale factor to apply to each feature.
    - Example: To double the range of all features, use `scale=2.0`.