Project Documentation: Baseline Model for Cross-Validation with Hyperparameter Tuning

Overview

This project provides a Python function, `Baseline`, to perform k-fold cross-validation and optional hyperparameter tuning on a machine learning estimator. The function evaluates model performance on different datasets using specified scoring metrics and saves the cross-validation results to a CSV file.

Key Components

1. **Libraries Used**:
   - `sklearn`: For model selection, scoring metrics, and dataset loading.
   - `pandas` and `numpy`: For data manipulation and storage.
2. **Function: `Baseline`**
   - **Purpose**: Performs k-fold cross-validation on the specified estimator. If a hyperparameter grid is provided, it applies grid search to identify the best parameters.
   - **Parameters**:
     - `estimator`: The machine learning model to train (e.g., `KNeighborsClassifier`).
     - `fold` (int): Number of folds for cross-validation (default is 10).
     - `X`, `y`: Feature matrix and target vector of the dataset.
     - `scoring` (str): Metric for model evaluation, e.g., `accuracy`, `f1_score`, `matthews_corrcoef`.
     - `param_grid` (dict, optional): Grid for hyperparameter tuning.
     - `dataset_name` (str): Name of the dataset for output file labeling.
   - **Returns**: Mean cross-validation score of the best model.
   - **Output**: A CSV file named `{dataset_name}_cross_validation_results.csv` containing the cross-validation scores and summary statistics.
3. **Datasets and Usage Examples**
   - The function is applied to several datasets (`iris`, `wine`, `vehicle`), with each dataset loaded and processed with k-nearest neighbors (`KNeighborsClassifier`).
   - For each dataset, the function performs a 10-fold cross-validation, optimizes the `n_neighbors` and `weights` parameters using `GridSearchCV`, and saves results to CSV files.
4. **Example of Running the Baseline Function**

```
from sklearn.datasets import load_iris
from sklearn.neighbors import KNeighborsClassifier

# Load the dataset
data = load_iris()
X, y = data.data, data.target

# Define model and parameter grid
estimator = KNeighborsClassifier()
```

```
param_grid = {'n_neighbors': [3, 5, 7], 'weights': ['uniform',
'distance']}

# Run the Baseline function
Baseline(estimator=estimator, fold=10, X=X, y=y, scoring='accuracy',
param_grid=param_grid, dataset_name="iris")
```

This saves the cross-validation results of the `iris` dataset to `iris_cross_validation_results.csv`.