explainer.py

Summary

Defines an abstract base class `ExplainerMixIn` for implementing explainability using SHAP values.

Dependencies

Standard Library

- abc

Other

- numpy
- shap (optional)

Description

This file defines an abstract base class `ExplainerMixIn` that provides a minimal interface for implementing explainability in machine learning models using SHAP (SHapley Additive exPlanations) values. The class is designed to be used as a mixin, allowing other classes to inherit its functionality.

The implementation includes a graceful fallback mechanism when the shap library is not available, ensuring that code using this module can still run without the explainability features.

The `ExplainerMixIn` class requires implementing classes to have an `explainer_` attribute of type `shap.Explainer` when the shap library is available. This attribute is expected to be initialized with a SHAP explainer object, which will be used to generate explanations for model predictions.

The class defines an abstract method `explain` that takes input data `X` and optional explainer parameters. This method is intended to return a dictionary containing the average response values and their respective SHAP values for each class in the case of a multi-class classifier.

Implementing classes must provide their own implementation of this method to generate explanations specific to their model architecture and use case.

By using this mixin, developers can easily add explainability features to their machine learning models, providing insights into how the model makes predictions and which features contribute most to those predictions.