#### ****1.**** 309\_preprocess.py ****(Preprocessing Script)****

This script performs the following preprocessing tasks:

**Feature Engineering**:

* + Computes Family\_Size by summing SibSp and Parch.

**Handling Missing Values**:

* + Fills missing Embarked values with the mode.
  + Uses a RandomForestRegressor to impute missing Age values.
  + Replaces infinite values with NaN and fills missing numerical values with the median.

**Categorical Encoding**:

* + One-hot encodes categorical features (Sex, Embarked).

**Creating Additional Features**:

* + Age\_Group: Categorizes passengers into different age bins.
  + Fare\_log: Applies a logarithmic transformation to the Fare column.

**Dropping Unnecessary Columns**:

* + Removes Cabin, Ticket, PassengerId, Name, etc.

**Feature Scaling**:

* + Scales numerical features (Age, Fare\_log, Family\_Size) using StandardScaler.

**Saving Output**:

* + Saves the preprocessed datasets as CSV files for further modeling.

#### ****2.**** preprocessing.yaml ****(Kubernetes Job Configuration)****

Defines a Kubernetes batch job to run the preprocessing pipeline as a containerized job.

* **Namespace**: aidp
* **Container Image**: triedlotsofnames/preprocess-container:latest
* **Persistent Volume**: Mounts data-pvc to /Data for storing preprocessed files.
* **Restart Policy**: Never (Ensures the job does not restart after completion).

#### ****3.**** Dockerfile ****(Container Setup)****

Defines a Docker container to execute the preprocessing script in a controlled environment.