Task: Spaceship Titanic Prediction

Problem Overview:

The Spaceship Titanic competition is a binary classification problem where the goal is to predict whether a passenger was transported to an alternate dimension based on given attributes.

Code Explanation:

Data Loading:

- The dataset is loaded using Pandas from `train.csv`.
- The first few rows, last few rows, and dataset summary statistics are displayed.
- `df.info()` is used to check for missing values and data types.
- Column names are extracted and printed.

Handling Missing Values:

- Numerical columns with missing values (`Age`, `RoomService`, `FoodCourt`, `Spa`, `ShoppingMall`, `VRDeck`) are filled with their respective mean values.
- These numerical columns are then converted to integers.
- Categorical columns (`HomePlanet`, `Destination`) are filled with the most frequent value (mode).
- Boolean columns (`VIP`, `CryoSleep`) are filled with `False` and converted to integers (0 or 1).
- Unnecessary columns like `PassengerId`, `Name`, and `Cabin` are dropped.

Encoding Categorical Data:

- The `HomePlanet` and `Destination` columns are encoded into numerical values using `LabelEncoder`.

Data Preprocessing for Testing Data:

- The same preprocessing steps applied to the training dataset are performed on the test dataset (`test.csv`).
- Numerical and categorical missing values are handled similarly.
- Unnecessary columns are dropped.

Model Training:

- Features (`X`) and target (`y`) are separated.

- The dataset is split into training and validation sets (80-20 split) using `train_test_split`.
- A `RandomForestClassifier` with 100 trees and a fixed random state is used for training.
- The model is trained on the training data (`X_train`, `y_train`).

Model Evaluation:

- Predictions are made on the validation set.
- The accuracy of the model is calculated using `accuracy_score`.

Predictions on Test Data:

- The test dataset is prepared by keeping only the required feature columns.
- Predictions are made on the test dataset.
- The results are stored in a submission file named `submission.csv`.

Output:

- The model outputs an accuracy score for the validation set.
- The submission file contains two columns: `PassengerId` and `Transported` (boolean values indicating whether the passenger was transported).

Conclusion:

This implementation of the Spaceship Titanic problem follows a structured approach for data cleaning, preprocessing, and model training using a Random Forest classifier. The pipeline ensures missing values are handled properly, categorical variables are encoded, and a submission file is generated for evaluation.