In [1]:
```
import pandas as pd
import numpy as np
import math
from sklearn.pipeline import Pipeline
from sklearn.base import ClassifierMixin, BaseEstimator, TransformerMixin
from sklearn.preprocessing import FunctionTransformer, MinMaxScaler, OneHotEncoder,
from sklearn.metrics import accuracy_score
from sklearn.compose import ColumnTransformer
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.model_selection import GridSearchCV
```

In [2]:
```
data = pd.read_csv('train.csv')
train = data
train_xs = train.drop(columns = "Transported")
train_ys = train['Transported']
test_xs = pd.read_csv('test.csv')
train_xs.dtypes
print(train_xs)
```

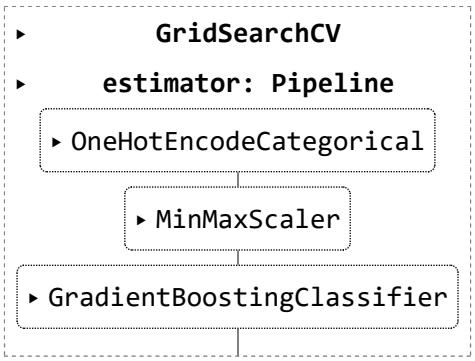
```
     PassengerId HomePlanet CryoSleep     Cabin    Destination   Age    VIP  \
0        0001_01     Europa     False     B/0/P    TRAPPIST-1e  39.0  False
1        0002_01      Earth     False     F/0/S    TRAPPIST-1e  24.0  False
2        0003_01     Europa     False     A/0/S    TRAPPIST-1e  58.0   True
3        0003_02     Europa     False     A/0/S    TRAPPIST-1e  33.0  False
4        0004_01      Earth     False     F/1/S    TRAPPIST-1e  16.0  False
...          ...        ...       ...       ...            ...   ...    ...
8688     9276_01     Europa     False    A/98/P    55 Cancri e  41.0   True
8689     9278_01      Earth      True   G/1499/S  PSO J318.5-22  18.0  False
8690     9279_01      Earth     False   G/1500/S    TRAPPIST-1e  26.0  False
8691     9280_01     Europa     False    E/608/S    55 Cancri e  32.0  False
8692     9280_02     Europa     False    E/608/S    TRAPPIST-1e  44.0  False

      RoomService  FoodCourt  ShoppingMall     Spa  VRDeck              Name
0             0.0        0.0           0.0     0.0     0.0    Maham Ofracculy
1           109.0        9.0          25.0   549.0    44.0       Juanna Vines
2            43.0     3576.0           0.0  6715.0    49.0      Altark Susent
3             0.0     1283.0         371.0  3329.0   193.0       Solam Susent
4           303.0       70.0         151.0   565.0     2.0  Willy Santantines
...           ...        ...           ...     ...     ...                ...
8688          0.0     6819.0           0.0  1643.0    74.0  Gravior Noxnuther
8689          0.0        0.0           0.0     0.0     0.0    Kurta Mondalley
8690          0.0        0.0        1872.0     1.0     0.0       Fayey Connon
8691          0.0     1049.0           0.0   353.0  3235.0   Celeon Hontichre
8692        126.0     4688.0           0.0     0.0    12.0   Propsh Hontichre

[8693 rows x 13 columns]
```

In [3]:
```
class OneHotEncodeCategorical(BaseEstimator, TransformerMixin):
    def fit(self, X, y=None):
        return self

    def transform(self, X):
        #print(temp)
        columns_to_drop = ["Name","PassengerId"]
```

```
        temp = X.drop(columns = columns_to_drop)
        temp.fillna(value=0,inplace=True)


####
        X_copy = temp.copy()
        cabin_info = temp['Cabin'].str.extract(r'(?P<Deck>[A-Za-z])/(?P<Number>\d+)
        #print(cabin_info['Deck'])
        #adding deck and number give a nan warning for test scores
        #X_copy = pd.concat([X_copy, cabin_info['Deck']], axis=1)
        #X_copy = pd.concat([X_copy, cabin_info['Number']], axis=1)

        X_copy = pd.concat([X_copy, cabin_info['Side']], axis=1)
        X_copy = X_copy.drop(columns="Cabin")
        temp = X_copy;
####



        #print(temp.)
        #print(temp.columns.tolist())
        #print(temp['Deck'])

        categorical_columns = temp.select_dtypes(include=['object']).columns
        X_encoded = pd.get_dummies(temp, columns=categorical_columns)
        print(X_encoded.columns.tolist())

        return X_encoded


gradientboosting_pipeline = Pipeline([
    ('ordinal_encoder', OneHotEncodeCategorical()),

    ('scaler',MinMaxScaler()),
    ('gradient_boosting', GradientBoostingClassifier())
])

gradientboosting_grid = {
    'gradient_boosting__subsample': [0.5,0.6,0.7,0.8,0.9,1],
    'gradient_boosting__learning_rate': [0.1,0.3,0.5],

    'gradient_boosting__n_estimators': [50,60,70,80,90,100],  #number of boosting s
    'gradient_boosting__max_depth': [3,4,5,6,7],  #limits number of nodes in tree
}
gradientboosting_search = GridSearchCV(gradientboosting_pipeline, gradientboosting_
gradientboosting_search.fit(train_xs, train_ys)
```

```
['Age', 'RoomService', 'FoodCourt', 'ShoppingMall', 'Spa', 'VRDeck', 'HomePlanet_0',
'HomePlanet_Earth', 'HomePlanet_Europa', 'HomePlanet_Mars', 'CryoSleep_False', 'Cryo
Sleep_True', 'Destination_0', 'Destination_55 Cancri e', 'Destination_PSO J318.5-22
', 'Destination_TRAPPIST-1e', 'VIP_False', 'VIP_True', 'Side_P', 'Side_S']
```

Out[3]:



In [4]:
```
gradientboosting_params = gradientboosting_search.best_params_
gradientboosting_score = gradientboosting_search.best_score_
print(f"Accuracy: {gradientboosting_score}")
print(f"Best params: {gradientboosting_params}\n")
```

```
Accuracy: 0.7962759836446882
Best params: {'gradient_boosting__learning_rate': 0.1, 'gradient_boosting__max_depth
': 4, 'gradient_boosting__n_estimators': 70, 'gradient_boosting__subsample': 0.9}
```

In [5]:
```
best_gradientboosting = gradientboosting_search.best_estimator_

predicted_values = best_gradientboosting.predict(test_xs)
passenger_ids = test_xs['PassengerId'].reset_index(drop=True)

result_df = pd.DataFrame({'PassengerId': passenger_ids, 'Transported': predicted_va

print(result_df)
result_df.to_csv('predicted_results.csv', index=False)
```

```
['Age', 'RoomService', 'FoodCourt', 'ShoppingMall', 'Spa', 'VRDeck', 'HomePlanet_0',
'HomePlanet_Earth', 'HomePlanet_Europa', 'HomePlanet_Mars', 'CryoSleep_False', 'Cryo
Sleep_True', 'Destination_0', 'Destination_55 Cancri e', 'Destination_PSO J318.5-22
', 'Destination_TRAPPIST-1e', 'VIP_False', 'VIP_True', 'Side_P', 'Side_S']
     PassengerId  Transported
0        0013_01         True
1        0018_01        False
2        0019_01         True
3        0021_01         True
4        0023_01         True
...          ...          ...
4272     9266_02         True
4273     9269_01        False
4274     9271_01         True
4275     9273_01         True
4276     9277_01         True

[4277 rows x 2 columns]
```

In [ ]: