Machine Learning

(1.) Data pre-processing : load a dataset, handle missing values, Encode categorical data, and normalize / standardize features.

```
Import pandas as pd
Import numpy as np
from sklearn.impute Import SimpleImputer
from sklearn.preprocessing Import OneHot
-Encoder, StandardScaler.
from sklearn.compose Import ColumnTransfo
-rmer
from sklearn.pipeline Import pipeline.
data = pd.Dataframe ({'Age': [25, np.nan,
 35, 40, 29], 'Salary': [50000, 6000, np.nan,
 80000, 52000], 'Department': ['Sales',
 'Engineering', 'HR', np.nan, 'Sales'],
'purchased': ['yes', 'No', 'yes', 'No', 'yes']
 })
print ("Original dataset:\n")
print (data)
X = data.drop ('purchased', axis=1)
y = data['purchased']
numerical_cols = X.select_dtypes (include=
 ['int64', 'float64']).columns.tolist()
categorical_cols = X.select_dtypes (include=
      = ['object']).colums.tolist()
numerical_pipeline = pipeline (steps=[
 ('imputer', SimpleImputer (strategy='mean')),
 ('scaler', StandardScaler ())
 ])
```

```
(imputation + One-hot Encoding)
categarical_pipeline = pipeline(steps=[
 ("Imputer", SimpleImputer(strategy
 = "most_frequent")),
('Encoder', OneHotEncoder(handle_unknown
   = "ignore"))
 ])
preprocessor = ColumnTransformer(trans
-formers=[
 ("num", numerical_pipeline, numerical
 -cols),
 ('cat', categarical_pipeline, categarical
 -cols)])
x_processed = preprocessor.fit_trans
-form(x)
print("\nprocessed features (after
   handling missing values, encoding,
 and scaling):\n")
print(x_processed.toarray() if
   hasattr(x_processed, "toarray")
else x_processed.
names encoded_feature_names.
preprocessor.named_transformers
 _['cat']['Encoder'].get_feature
_name_out.(categorical_cols)
all.features_names = numerical_cols
+ Encoded_features_nams.tolist()
Print("\n processed features Names:")
print(all_feature_names)
```

OUTPUT :

Original Dataset.

```
      Age.    Salary     department    purchased

0     25.0    50000      Sales         Yes.

1     Nan.    60000.0    Engineering   No.

3.    35.0    Nan        HR            Yes.
```

processed feature Names.

```
['Age', 'Salary', 'Department_Engineering,
 'department_HR', 'Department_Sales'].
```

processed features (after handling missing values, encoding & scaling):

```
[[-1.41421356   -1.18321596   1.   0.   0.]
 [  0.            0.          0.   1.   0.]
 [0.70710678    -0.50709255   0.   0.   1.]
 [1.41421356     1.52127766   1.   0.   0.]
 [-0.70710678   -0.16903085   1.   0.   0.]]
```