DATA PREPROCESSING

In [42]:
```
* Import the libraries
* Importing the Dataset
* Checking for null values
* Data Visualization
* Outlier Detection
* Splitting Dependent and Independent variables
* Perform Encoding
* Feature Scaling
* Splitting Data into Train and Test
```

```
  Cell In[42], line 1
    * Import the libraries
            ^
SyntaxError: invalid syntax
```

Import the Libraries

In [ ]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

Importing the Dataset

In [ ]:
```
df=pd.read_csv("Titanic-Dataset.csv")
```

In [ ]:
```
df.head()
```

In [43]:
```
df.describe()
```

Out[43]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

In [44]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  891 non-null    int64  
 1   Survived     891 non-null    int64  
 2   Pclass       891 non-null    int64  
 3   Name         891 non-null    object 
 4   Sex          891 non-null    object 
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64  
 7   Parch        891 non-null    int64  
 8   Ticket       891 non-null    object 
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object 
 11  Embarked     889 non-null    object 
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

Checking For Null Values

In [45]:
```
df.isnull().any()
```

Out[45]:
```
PassengerId    False
Survived       False
Pclass         False
Name           False
Sex            False
Age             True
SibSp          False
Parch          False
Ticket         False
Fare           False
Cabin           True
Embarked        True
dtype: bool
```

In [46]:
```
df.isnull().sum()
```

Out[46]:
```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

In [47]:
```
df.Age.nunique()
```

Out[47]: 88

In [48]:
```
df.Age.unique()
```

Out[48]:
```
array([22.  , 38.  , 26.  , 35.  ,   nan, 54.  ,  2.  , 27.  , 14.  ,
        4.  , 58.  , 20.  , 39.  , 55.  , 31.  , 34.  , 15.  , 28.  ,
        8.  , 19.  , 40.  , 66.  , 42.  , 21.  , 18.  ,  3.  ,  7.  ,
       49.  , 29.  , 65.  , 28.5 ,  5.  , 11.  , 45.  , 17.  , 32.  ,
       16.  , 25.  ,  0.83, 30.  , 33.  , 23.  , 24.  , 46.  , 59.  ,
       71.  , 37.  , 47.  , 14.5 , 70.5 , 32.5 , 12.  ,  9.  , 36.5 ,
       51.  , 55.5 , 40.5 , 44.  ,  1.  , 61.  , 56.  , 50.  , 36.  ,
       45.5 , 20.5 , 62.  , 41.  , 52.  , 63.  , 23.5 ,  0.92, 43.  ,
       60.  , 10.  , 64.  , 13.  , 48.  ,  0.75, 53.  , 57.  , 80.  ,
       70.  , 24.5 ,  6.  ,  0.67, 30.5 ,  0.42, 34.5 , 74.  ])
```

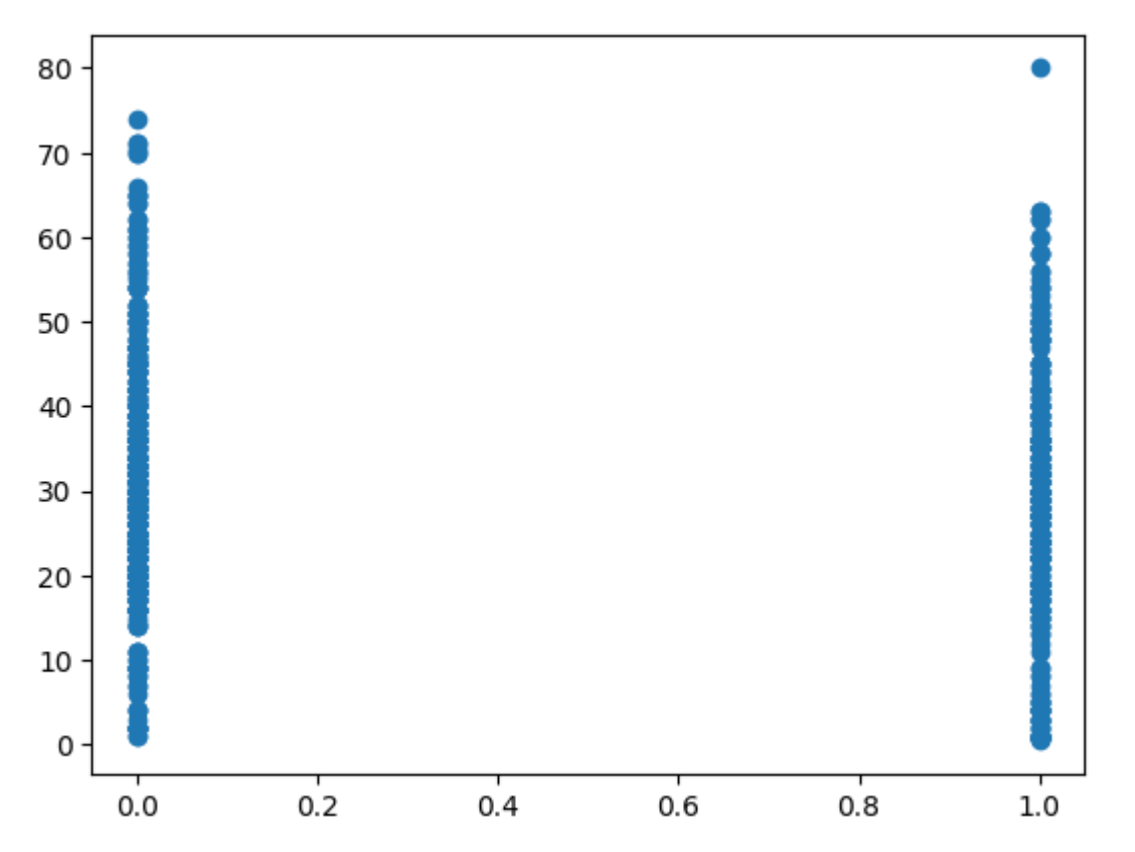
In [49]:
```
df.Age.value_counts()
```

Out[49]:
```
Age
24.00    30
22.00    27
18.00    26
19.00    25
28.00    25
         ..
36.50     1
55.50     1
0.92      1
23.50     1
74.00     1
Name: count, Length: 88, dtype: int64
```

DATA VISUALIZATION

In [50]:
```
plt.scatter(df["Survived"],df["Age"])
```

Out[50]: <matplotlib.collections.PathCollection at 0x2ae4d0663d0>



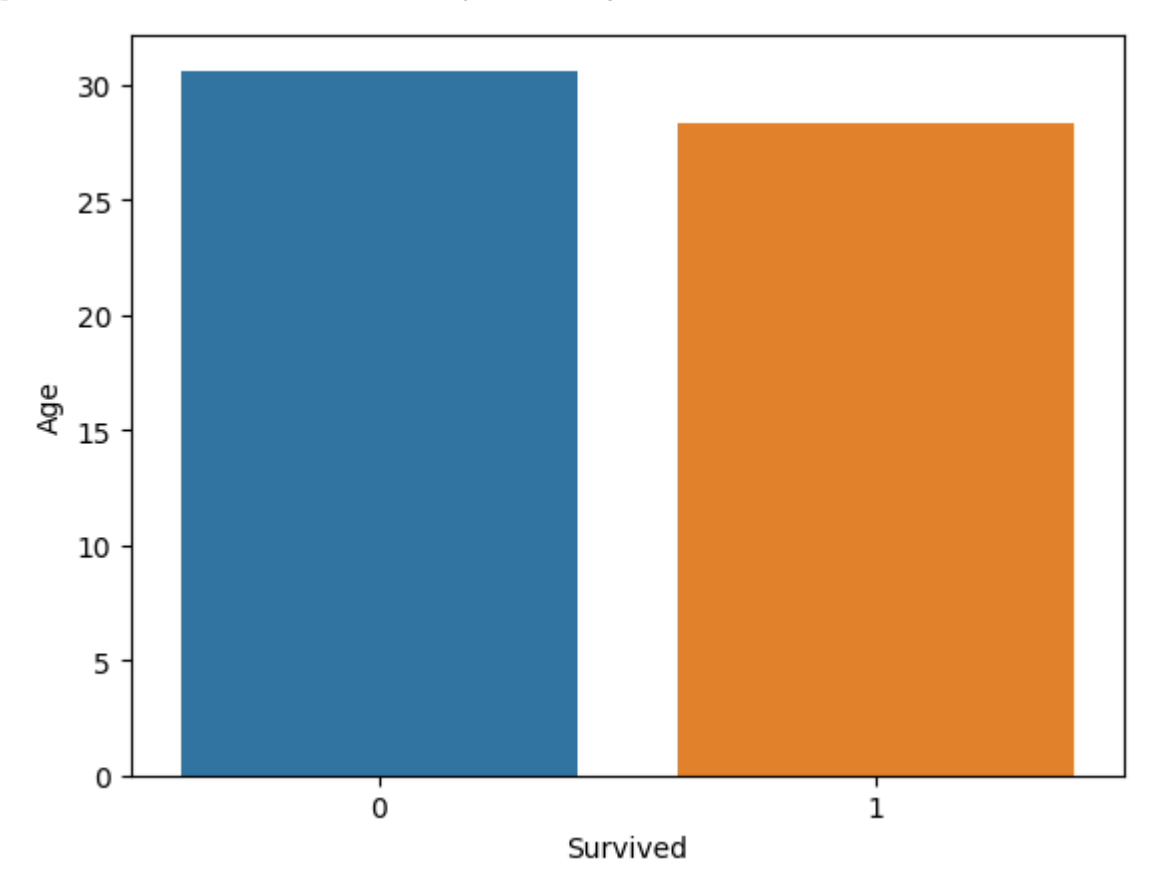
In [51]:
```
sns.barplot(x=df["Survived"],y=df["Age"],ci=0)
```

```
C:\Users\sgowt\AppData\Local\Temp\ipykernel_16732\332883349.py:1: FutureWarning: 

The `ci` parameter is deprecated. Use `errorbar=('ci', 0)` for the same effect.

  sns.barplot(x=df["Survived"],y=df["Age"],ci=0)
C:\Users\sgowt\AppData\Local\Programs\Python\Python311\Lib\site-packages\seaborn\_oldcore.py:1498: FutureWarning: is_categorical_dtype is deprecated and will be removed in a future version. Use isinstance(dtype, CategoricalDtype) inste
ad
  if pd.api.types.is_categorical_dtype(vector):
C:\Users\sgowt\AppData\Local\Programs\Python\Python311\Lib\site-packages\seaborn\_oldcore.py:1498: FutureWarning: is_categorical_dtype is deprecated and will be removed in a future version. Use isinstance(dtype, CategoricalDtype) inste
ad
  if pd.api.types.is_categorical_dtype(vector):
C:\Users\sgowt\AppData\Local\Programs\Python\Python311\Lib\site-packages\seaborn\_oldcore.py:1498: FutureWarning: is_categorical_dtype is deprecated and will be removed in a future version. Use isinstance(dtype, CategoricalDtype) inste
ad
  if pd.api.types.is_categorical_dtype(vector):
```

Out[51]: <Axes: xlabel='Survived', ylabel='Age'>



Outlier Detection

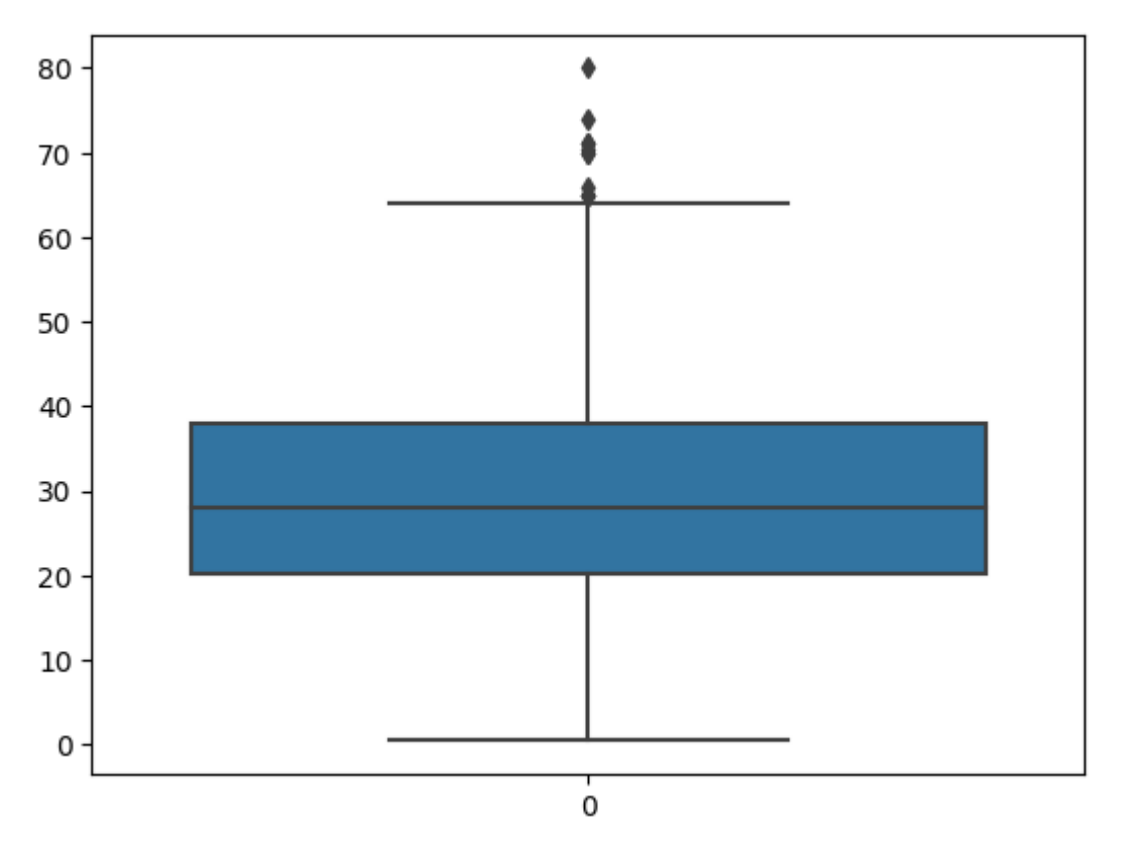
In [52]:
```
df.head()
```

Out[52]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

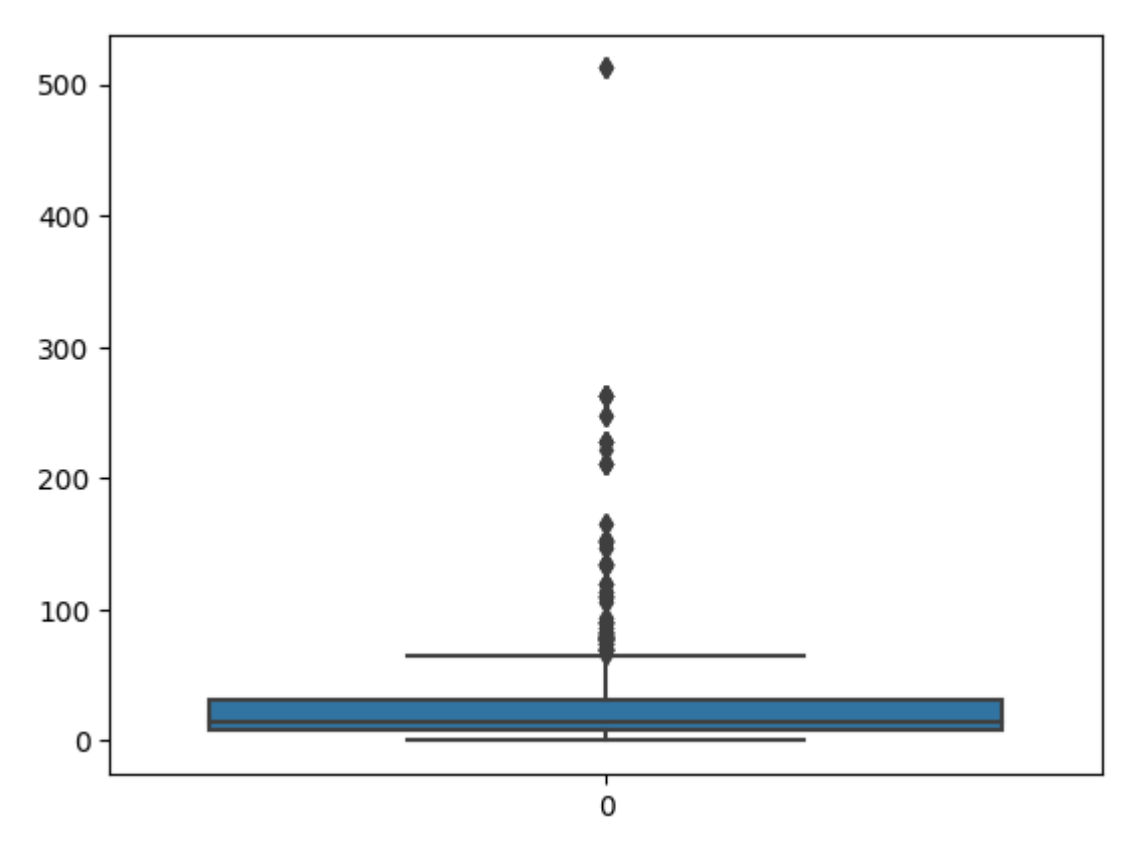
In [53]:
```
sns.boxplot(df["Age"])
```

Out[53]: <Axes: >



In [54]:
```
sns.boxplot(df["Fare"])
```

Out[54]: <Axes: >



Splitting Dependent and Independent Variables

In [55]:
```
df.head()
```

Out[55]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

In [56]:
```
X=df.drop(columns=["Age"],axis=1)
X.head()
```

Out[56]:

	PassengerId	Survived	Pclass	Name	Sex	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	0	0	373450	8.0500	NaN	S

In [57]:
```
X.shape
```

Out[57]: (891, 11)

In [58]:
```
type(X)
```

Out[58]: pandas.core.frame.DataFrame

In [59]:
```
y=df["Age"]
y.head()
```

Out[59]:
```
0    22.0
1    38.0
2    26.0
3    35.0
4    35.0
Name: Age, dtype: float64
```

ENCODING

In [60]:
```
X.head()
```

Out[60]:

	PassengerId	Survived	Pclass	Name	Sex	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	0	0	373450	8.0500	NaN	S

In [75]:
```
import sklearn
```

In [76]:
```
from sklearn.preprocessing import LabelEncoder
le=LabelEncoder()
```

In [84]:
```
X["Age"]=le.fit_transform(X["Age"])
```

```
---------------------------------------------------------------------------
KeyError                                  Traceback (most recent call last)
File ~\AppData\Local\Programs\Python\Python311\Lib\site-packages\pandas\core\indexes\base.py:3790, in Index.get_loc(self, key)
   3789 try:
-> 3790     return self._engine.get_loc(casted_key)
   3791 except KeyError as err:

File index.pyx:152, in pandas._libs.index.IndexEngine.get_loc()

File index.pyx:181, in pandas._libs.index.IndexEngine.get_loc()

File pandas\_libs\hashtable_class_helper.pxi:7080, in pandas._libs.hashtable.PyObjectHashTable.get_item()

File pandas\_libs\hashtable_class_helper.pxi:7088, in pandas._libs.hashtable.PyObjectHashTable.get_item()

KeyError: 'Age'

The above exception was the direct cause of the following exception:

KeyError                                  Traceback (most recent call last)
Cell In[84], line 1
----> 1 X["Age"]=le.fit_transform(X["Age"])

File ~\AppData\Local\Programs\Python\Python311\Lib\site-packages\pandas\core\frame.py:3896, in DataFrame.__getitem__(self, key)
   3894 if self.columns.nlevels > 1:
   3895     return self._getitem_multilevel(key)
-> 3896 indexer = self.columns.get_loc(key)
   3897 if is_integer(indexer):
   3898     indexer = [indexer]

File ~\AppData\Local\Programs\Python\Python311\Lib\site-packages\pandas\core\indexes\base.py:3797, in Index.get_loc(self, key)
   3792     if isinstance(casted_key, slice) or (
   3793         isinstance(casted_key, abc.Iterable)
   3794         and any(isinstance(x, slice) for x in casted_key)
   3795     ):
   3796         raise InvalidIndexError(key)
-> 3797     raise KeyError(key) from err
   3798 except TypeError:
   3799     # If we have a listlike key, _check_indexing_error will raise
   3800     #  InvalidIndexError. Otherwise we fall through and re-raise
   3801     #  the TypeError.
   3802     self._check_indexing_error(key)

KeyError: 'Age'
```

In [ ]:
```
X.head()
```

print(le.classes_)

In [83]:

```
---------------------------------------------------------------------------
KeyError                                  Traceback (most recent call last)
Cell In[83], line 1
----> 1 df = pd.get_dummies(df,columns=['Age','Fare'],drop_first=True)

File ~\AppData\Local\Programs\Python\Python311\Lib\site-packages\pandas\core\reshape\encoding.py:164, in get_dummies(data, prefix, prefix_sep, dummy_na, columns, sparse, drop_first, dtype)
    162     raise TypeError("Input must be a list-like for parameter `columns`")
    163 else:
--> 164     data_to_encode = data[columns]
    166 # validate prefixes and separator to avoid silently dropping cols
    167 def check_len(item, name: str):

File ~\AppData\Local\Programs\Python\Python311\Lib\site-packages\pandas\core\frame.py:3902, in DataFrame.__getitem__(self, key)
   3900     if is_iterator(key):
   3901         key = list(key)
-> 3902     indexer = self.columns._get_indexer_strict(key, "columns")[1]
   3904 # take() does not accept boolean indexers
   3905 if getattr(indexer, "dtype", None) == bool:

File ~\AppData\Local\Programs\Python\Python311\Lib\site-packages\pandas\core\indexes\base.py:6114, in Index._get_indexer_strict(self, key, axis_name)
   6111 else:
   6112     keyarr, indexer, new_indexer = self._reindex_non_unique(keyarr)
-> 6114 self._raise_if_missing(keyarr, indexer, axis_name)
   6116 keyarr = self.take(indexer)
   6117 if isinstance(key, Index):
   6118     # GH 42790 - Preserve name from an Index

File ~\AppData\Local\Programs\Python\Python311\Lib\site-packages\pandas\core\indexes\base.py:6175, in Index._raise_if_missing(self, key, indexer, axis_name)
   6173     if use_interval_msg:
   6174         key = list(key)
-> 6175     raise KeyError(f"None of [{key}] are in the [{axis_name}]")
   6177 not_found = list(ensure_index(key)[missing_mask.nonzero()[0]].unique())
   6178 raise KeyError(f"{not_found} not in index")

KeyError: "None of [Index(['Age', 'Fare'], dtype='object')] are in the [columns]"
```

FEATURE SCALING

In [ ]:
```
from sklearn.preprocessing import MinMaxScaler
ms=MinMaxScaler()
```

In [ ]:
```
X_Scaled=ms.fit_transform(X)
```

In [ ]:
```
X_Scaled=pd.Dataframe(ms.fit_transform(X),columns=X.columns)
```

In [ ]:
```
X_Scaled.head()
```

Train Test Split

In [80]:
```
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test = train_test_split(X_scaled,y_test_size = 0.2,random_state=0)
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[80], line 2
      1 from sklearn.model_selection import train_test_split
----> 2 x_train,x_test,y_train,y_test = train_test_split(X_scaled,y_test_size = 0.2,random_state=0)

NameError: name 'X_scaled' is not defined
```

In [78]:
```
print(x_train.shape,x_test.shape,y_train.shape,y_test.shape)
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[78], line 1
----> 1 print(x_train.shape,x_test.shape,y_train.shape,y_test.shape)
```

```
NameError: name 'x_train' is not defined
```

In [ ]: