NAME: ADITYA SONI

ROLL NO.: 65

PID: 246052

BATCH: 3

TITLE: IMPLEMENTATION OF AN END-TO-END MACHINE LEARNING DATA PIPELINE

In [3]:
```
%pip install numpy pandas matplotlib seaborn scikit-learn
```

```
Defaulting to user installation because normal site-packages is not writeable
Requirement already satisfied: numpy in c:\users\hp\appdata\roaming\python\python39
\site-packages (2.0.2)
Requirement already satisfied: pandas in c:\users\hp\appdata\roaming\python\python39
\site-packages (2.3.3)
Requirement already satisfied: matplotlib in c:\users\hp\appdata\roaming\python\pyth
on39\site-packages (3.9.4)
Requirement already satisfied: seaborn in c:\users\hp\appdata\roaming\python\python3
9\site-packages (0.13.2)
Collecting scikit-learn
  Downloading scikit_learn-1.6.1-cp39-cp39-win_amd64.whl.metadata (15 kB)
Requirement already satisfied: python-dateutil>=2.8.2 in c:\users\hp\appdata\roaming
\python\python39\site-packages (from pandas) (2.9.0.post0)
Requirement already satisfied: pytz>=2020.1 in c:\users\hp\appdata\roaming\python\py
thon39\site-packages (from pandas) (2025.2)
Requirement already satisfied: tzdata>=2022.7 in c:\users\hp\appdata\roaming\python
\python39\site-packages (from pandas) (2025.3)
Requirement already satisfied: contourpy>=1.0.1 in c:\users\hp\appdata\roaming\pytho
n\python39\site-packages (from matplotlib) (1.3.0)
Requirement already satisfied: cycler>=0.10 in c:\users\hp\appdata\roaming\python\py
thon39\site-packages (from matplotlib) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in c:\users\hp\appdata\roaming\pyth
on\python39\site-packages (from matplotlib) (4.60.1)
Requirement already satisfied: kiwisolver>=1.3.1 in c:\users\hp\appdata\roaming\pyth
on\python39\site-packages (from matplotlib) (1.4.7)
Requirement already satisfied: packaging>=20.0 in c:\users\hp\appdata\roaming\python
\python39\site-packages (from matplotlib) (25.0)
Requirement already satisfied: pillow>=8 in c:\users\hp\appdata\roaming\python\pytho
n39\site-packages (from matplotlib) (11.3.0)
Requirement already satisfied: pyparsing>=2.3.1 in c:\users\hp\appdata\roaming\pytho
n\python39\site-packages (from matplotlib) (3.2.5)
Requirement already satisfied: importlib-resources>=3.2.0 in c:\users\hp\appdata\roa
ming\python\python39\site-packages (from matplotlib) (6.5.2)
Collecting scipy>=1.6.0 (from scikit-learn)
  Downloading scipy-1.13.1-cp39-cp39-win_amd64.whl.metadata (60 kB)
Collecting joblib>=1.2.0 (from scikit-learn)
  Downloading joblib-1.5.3-py3-none-any.whl.metadata (5.5 kB)
Collecting threadpoolctl>=3.1.0 (from scikit-learn)
  Downloading threadpoolctl-3.6.0-py3-none-any.whl.metadata (13 kB)
Requirement already satisfied: zipp>=3.1.0 in c:\users\hp\appdata\roaming\python\pyt
hon39\site-packages (from importlib-resources>=3.2.0->matplotlib) (3.23.0)
Requirement already satisfied: six>=1.5 in c:\users\hp\appdata\roaming\python\python
39\site-packages (from python-dateutil>=2.8.2->pandas) (1.17.0)
Downloading scikit_learn-1.6.1-cp39-cp39-win_amd64.whl (11.2 MB)
   ---------------------------------------- 0.0/11.2 MB ? eta -:--:--
   -------------- ------------------------ 4.2/11.2 MB 27.9 MB/s eta 0:00:01
   ----------------------- ---------------- 6.8/11.2 MB 17.5 MB/s eta 0:00:01
   -------------------------------- ------- 9.2/11.2 MB 15.4 MB/s eta 0:00:01
   ---------------------------------------  11.0/11.2 MB 13.8 MB/s eta 0:00:01
   ---------------------------------------- 11.2/11.2 MB 13.4 MB/s  0:00:00
Downloading joblib-1.5.3-py3-none-any.whl (309 kB)
Downloading scipy-1.13.1-cp39-cp39-win_amd64.whl (46.2 MB)
   ---------------------------------------- 0.0/46.2 MB ? eta -:--:--
   - -------------------------------------- 2.1/46.2 MB 9.0 MB/s eta 0:00:05
   ---- ----------------------------------- 5.5/46.2 MB 12.4 MB/s eta 0:00:04
   -------- ------------------------------- 10.2/46.2 MB 15.9 MB/s eta 0:00:03
```

```
   ------------ --------------------------- 13.9/46.2 MB 16.5 MB/s eta 0:00:02
   --------------- ------------------------ 18.4/46.2 MB 17.3 MB/s eta 0:00:02
   ------------------- -------------------- 22.5/46.2 MB 17.6 MB/s eta 0:00:02
   ------------------------- -------------- 29.6/46.2 MB 19.8 MB/s eta 0:00:01
   ----------------------------- ---------- 34.1/46.2 MB 20.2 MB/s eta 0:00:01
   -------------------------------- ------- 38.0/46.2 MB 20.0 MB/s eta 0:00:01
   ------------------------------------ --- 41.9/46.2 MB 19.8 MB/s eta 0:00:01
   ---------------------------------------  46.1/46.2 MB 20.0 MB/s eta 0:00:01
   ---------------------------------------- 46.2/46.2 MB 19.4 MB/s  0:00:02
Downloading threadpoolctl-3.6.0-py3-none-any.whl (18 kB)
Installing collected packages: threadpoolctl, scipy, joblib, scikit-learn

   ---------- ----------------------------- 1/4 [scipy]
   ---------- ----------------------------- 1/4 [scipy]
   ---------- ----------------------------- 1/4 [scipy]
   ---------- ----------------------------- 1/4 [scipy]
   ---------- ----------------------------- 1/4 [scipy]
   ---------- ----------------------------- 1/4 [scipy]
   ---------- ----------------------------- 1/4 [scipy]
   ---------- ----------------------------- 1/4 [scipy]
   ---------- ----------------------------- 1/4 [scipy]
   ---------- ----------------------------- 1/4 [scipy]
   ---------- ----------------------------- 1/4 [scipy]
   ---------- ----------------------------- 1/4 [scipy]
   ---------- ----------------------------- 1/4 [scipy]
   ---------- ----------------------------- 1/4 [scipy]
   ---------- ----------------------------- 1/4 [scipy]
   ---------- ----------------------------- 1/4 [scipy]
   ---------- ----------------------------- 1/4 [scipy]
   ---------- ----------------------------- 1/4 [scipy]
   ---------- ----------------------------- 1/4 [scipy]
   ---------- ----------------------------- 1/4 [scipy]
   ---------- ----------------------------- 1/4 [scipy]
   ---------- ----------------------------- 1/4 [scipy]
   ---------- ----------------------------- 1/4 [scipy]
   ---------- ----------------------------- 1/4 [scipy]
   ---------- ----------------------------- 1/4 [scipy]
   ---------- ----------------------------- 1/4 [scipy]
   -------------------- ------------------- 2/4 [joblib]
   -------------------- ------------------- 2/4 [joblib]
   ------------------------------ --------- 3/4 [scikit-learn]
   ------------------------------ --------- 3/4 [scikit-learn]
   ------------------------------ --------- 3/4 [scikit-learn]
   ------------------------------ --------- 3/4 [scikit-learn]
   ------------------------------ --------- 3/4 [scikit-learn]
   ------------------------------ --------- 3/4 [scikit-learn]
   ------------------------------ --------- 3/4 [scikit-learn]
   ------------------------------ --------- 3/4 [scikit-learn]
   ------------------------------ --------- 3/4 [scikit-learn]
   ------------------------------ --------- 3/4 [scikit-learn]
   ------------------------------ --------- 3/4 [scikit-learn]
   ------------------------------ --------- 3/4 [scikit-learn]
   ------------------------------ --------- 3/4 [scikit-learn]
   ------------------------------ --------- 3/4 [scikit-learn]
   ------------------------------ --------- 3/4 [scikit-learn]
   ------------------------------ --------- 3/4 [scikit-learn]
```

```
   ------------------------------ --------- 3/4 [scikit-learn]
   ------------------------------ --------- 3/4 [scikit-learn]
   ------------------------------ --------- 3/4 [scikit-learn]
   ------------------------------ --------- 3/4 [scikit-learn]
   ------------------------------ --------- 3/4 [scikit-learn]
   ------------------------------ --------- 3/4 [scikit-learn]
   ------------------------------ --------- 3/4 [scikit-learn]
   ------------------------------ --------- 3/4 [scikit-learn]
   ------------------------------ --------- 3/4 [scikit-learn]
   ---------------------------------------- 4/4 [scikit-learn]

Successfully installed joblib-1.5.3 scikit-learn-1.6.1 scipy-1.13.1 threadpoolctl-3.
6.0
Note: you may need to restart the kernel to use updated packages.
```

In [4]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

In [6]:
```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.preprocessing import LabelEncoder
from sklearn.metrics import accuracy_score
```

In [7]:
```
import seaborn as sns
titanic_data = sns.load_dataset('titanic')
```

In [8]:
```
print(titanic_data.shape)
```

```
(891, 15)
```

In [9]:
```
print(titanic_data.columns)
```

```
Index(['survived', 'pclass', 'sex', 'age', 'sibsp', 'parch', 'fare',
       'embarked', 'class', 'who', 'adult_male', 'deck', 'embark_town',
       'alive', 'alone'],
      dtype='object')
```

In [10]:
```
print(titanic_data.head())
```

```
   survived  pclass     sex   age  sibsp  parch     fare embarked  class  \
0         0       3    male  22.0      1      0   7.2500        S  Third
1         1       1  female  38.0      1      0  71.2833        C  First
2         1       3  female  26.0      0      0   7.9250        S  Third
3         1       1  female  35.0      1      0  53.1000        S  First
4         0       3    male  35.0      0      0   8.0500        S  Third

     who  adult_male deck  embark_town alive  alone
0    man        True  NaN  Southampton    no  False
1  woman       False    C    Cherbourg   yes  False
2  woman       False  NaN  Southampton   yes   True
3  woman       False    C  Southampton   yes  False
4    man        True  NaN  Southampton    no   True
```

In [11]:
```
print(titanic_data.tail())
```

```
     survived  pclass     sex   age  sibsp  parch   fare embarked   class  \
886         0       2    male  27.0      0      0  13.00        S  Second
887         1       1  female  19.0      0      0  30.00        S   First
888         0       3  female   NaN      1      2  23.45        S   Third
889         1       1    male  26.0      0      0  30.00        C   First
890         0       3    male  32.0      0      0   7.75        Q   Third

       who  adult_male deck  embark_town alive  alone
886    man        True  NaN  Southampton    no   True
887  woman       False    B  Southampton   yes   True
888  woman       False  NaN  Southampton    no  False
889    man        True    C    Cherbourg   yes   True
890    man        True  NaN   Queenstown    no   True
```

In [12]:
```
print(titanic_data.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 15 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   survived     891 non-null    int64
 1   pclass       891 non-null    int64
 2   sex          891 non-null    object
 3   age          714 non-null    float64
 4   sibsp        891 non-null    int64
 5   parch        891 non-null    int64
 6   fare         891 non-null    float64
 7   embarked     889 non-null    object
 8   class        891 non-null    category
 9   who          891 non-null    object
 10  adult_male   891 non-null    bool
 11  deck         203 non-null    category
 12  embark_town  889 non-null    object
 13  alive        891 non-null    object
 14  alone        891 non-null    bool
dtypes: bool(2), category(2), float64(2), int64(4), object(5)
memory usage: 80.7+ KB
None
```

In [13]:
```
print(titanic_data.describe())
```

```
        survived      pclass         age       sibsp       parch        fare
count  891.000000  891.000000  714.000000  891.000000  891.000000  891.000000
mean     0.383838    2.308642   29.699118    0.523008    0.381594   32.204208
std      0.486592    0.836071   14.526497    1.102743    0.806057   49.693429
min      0.000000    1.000000    0.420000    0.000000    0.000000    0.000000
25%      0.000000    2.000000   20.125000    0.000000    0.000000    7.910400
50%      0.000000    3.000000   28.000000    0.000000    0.000000   14.454200
75%      1.000000    3.000000   38.000000    1.000000    0.000000   31.000000
max      1.000000    3.000000   80.000000    8.000000    6.000000  512.329200
```

In [14]:
```
missing_values = titanic_data.isnull().sum()
print(missing_values)
```

```
survived         0
pclass           0
sex              0
age            177
sibsp            0
parch            0
fare             0
embarked         2
class            0
who              0
adult_male       0
deck           688
embark_town      2
alive            0
alone            0
dtype: int64
```

In [16]:
```
new_titanic_df = titanic_data.drop(columns = ['deck'])
```

In [40]:
```
new_titanic_df['age'] = (new_titanic_df['age'] - new_titanic_df['age'].min()) / (ne
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[40], line 1
----> 1 new_titanic_df['age'] = (new_titanic_df['age'] - titanic_df['age'].min()) /
(titanic_df['age'].max() - titanic_df['age'].min())

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

In [19]:
```
missing_values = new_titanic_df.isnull().sum()
print(missing_values)
```

```
survived       0
pclass         0
sex            0
age            0
sibsp          0
parch          0
fare           0
embarked       2
class          0
who            0
adult_male     0
embark_town    2
alive          0
alone          0
dtype: int64
```

In [20]:
```
data = new_titanic_df
data['embark_town'].dtype
data['embark_town'].unique()
data['embark_town'].fillna(data['embark_town'].mode()[0] , inplace=True)
data.isnull().sum()
```

```
C:\Users\HP\AppData\Local\Temp\ipykernel_7680\1968151471.py:4: FutureWarning: A valu
e is trying to be set on a copy of a DataFrame or Series through chained assignment
using an inplace method.
The behavior will change in pandas 3.0. This inplace method will never work because
the intermediate object on which we are setting values always behaves as a copy.

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method
({col: value}, inplace=True)' or df[col] = df[col].method(value) instead, to perform
the operation inplace on the original object.


  data['embark_town'].fillna(data['embark_town'].mode()[0] , inplace=True)
```

Out[20]:
```
survived       0
pclass         0
sex            0
age            0
sibsp          0
parch          0
fare           0
embarked       2
class          0
who            0
adult_male     0
embark_town    0
alive          0
alone          0
dtype: int64
```

In [21]:
```
data = new_titanic_df
data['embarked'].dtype
data['embarked'].unique()
data['embarked'].fillna(data['embarked'].mode()[0] , inplace=True)
data.isnull().sum()
```

```
C:\Users\HP\AppData\Local\Temp\ipykernel_7680\2633234362.py:4: FutureWarning: A valu
e is trying to be set on a copy of a DataFrame or Series through chained assignment
using an inplace method.
The behavior will change in pandas 3.0. This inplace method will never work because
the intermediate object on which we are setting values always behaves as a copy.

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method
({col: value}, inplace=True)' or df[col] = df[col].method(value) instead, to perform
the operation inplace on the original object.


  data['embarked'].fillna(data['embarked'].mode()[0] , inplace=True)
```

Out[21]:
```
survived       0
pclass         0
sex            0
age            0
sibsp          0
parch          0
fare           0
embarked       0
class          0
who            0
adult_male     0
embark_town    0
alive          0
alone          0
dtype: int64
```

In [22]:
```
le = LabelEncoder()
data['sex'] = le.fit_transform(data['sex'])
data['embarked'] = le.fit_transform(data['embarked'])
```

In [23]:
```
data.head()
```

Out[23]:

	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male
0	0	3	1	22.0	1	0	7.2500	2	Third	man	True
1	1	1	0	38.0	1	0	71.2833	0	First	woman	False
2	1	3	0	26.0	0	0	7.9250	2	Third	woman	False
3	1	1	0	35.0	1	0	53.1000	2	First	woman	False
4	0	3	1	35.0	0	0	8.0500	2	Third	man	True

In [24]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 14 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   survived     891 non-null    int64
 1   pclass       891 non-null    int64
 2   sex          891 non-null    int64
 3   age          891 non-null    float64
 4   sibsp        891 non-null    int64
 5   parch        891 non-null    int64
 6   fare         891 non-null    float64
 7   embarked     891 non-null    int64
 8   class        891 non-null    category
 9   who          891 non-null    object
 10  adult_male   891 non-null    bool
 11  embark_town  891 non-null    object
 12  alive        891 non-null    object
 13  alone        891 non-null    bool
dtypes: bool(2), category(1), float64(2), int64(6), object(3)
memory usage: 79.4+ KB
```

In [25]:
```
data = data[['pclass', 'sex' , 'age' , 'fare' , 'embarked' , 'survived' ]]
X = data[['pclass', 'sex' , 'age' , 'fare' , 'embarked' ]]
Y = data['survived']
```

In [26]:
```
X_train , X_test , Y_train , Y_test = train_test_split(X,Y,test_size = 0.3 , random
```

In [29]:
```
model = LogisticRegression(max_iter = 1000)
model.fit(X_train,Y_train)
```

Out[29]:

LogisticRegression

LogisticRegression(max_iter=1000)

In [30]:
```
Y_pred = model.predict(X_test)
```

In [31]:
```
accuracy = accuracy_score(Y_test,Y_pred)
print("Accuracy:",accuracy)
```

```
Accuracy: 0.7947761194029851
```

In [32]:
```
new_passenger = pd.DataFrame({
    'pclass' : [3],
    'sex' : ['male'],
    'age' : [28],
    'fare' : [7.25],
    'embarked' : ['S']
})
```

In [33]:
```
new_passenger_encoded = pd.get_dummies(new_passenger)
new_passenger_encoded = new_passenger_encoded.reindex(columns=X.columns , fill_valu
```

In [34]:
```
prediction = model.predict(new_passenger_encoded)
print("Survived" if prediction[0] == 1 else "Did not survive")
```

```
Survived
```

In [35]:
```
new_passenger = pd.DataFrame({
    'pclass' : [1,3,2],
    'sex' : ['female','male','female'],
    'age' : [38,45,14],
    'fare' : [80.0,8.05,20.0],
    'embarked' : ['C','S','Q']
})
```

In [36]:
```
new_passenger_encoded = pd.get_dummies(new_passenger)
new_passenger_encoded = new_passenger_encoded.reindex(columns=X.columns , fill_valu
```

In [37]:
```
prediction = model.predict(new_passenger_encoded)
print("Survived" if prediction[0] == 1 else "Did not survive")
```

```
Survived
```

In [38]:
```
predictions = model.predict(new_passenger_encoded)
for i,pred in enumerate(predictions):
    print(f"Passenger {i+1}:",
          "Survived" if pred == 1 else "Did not survive")
```

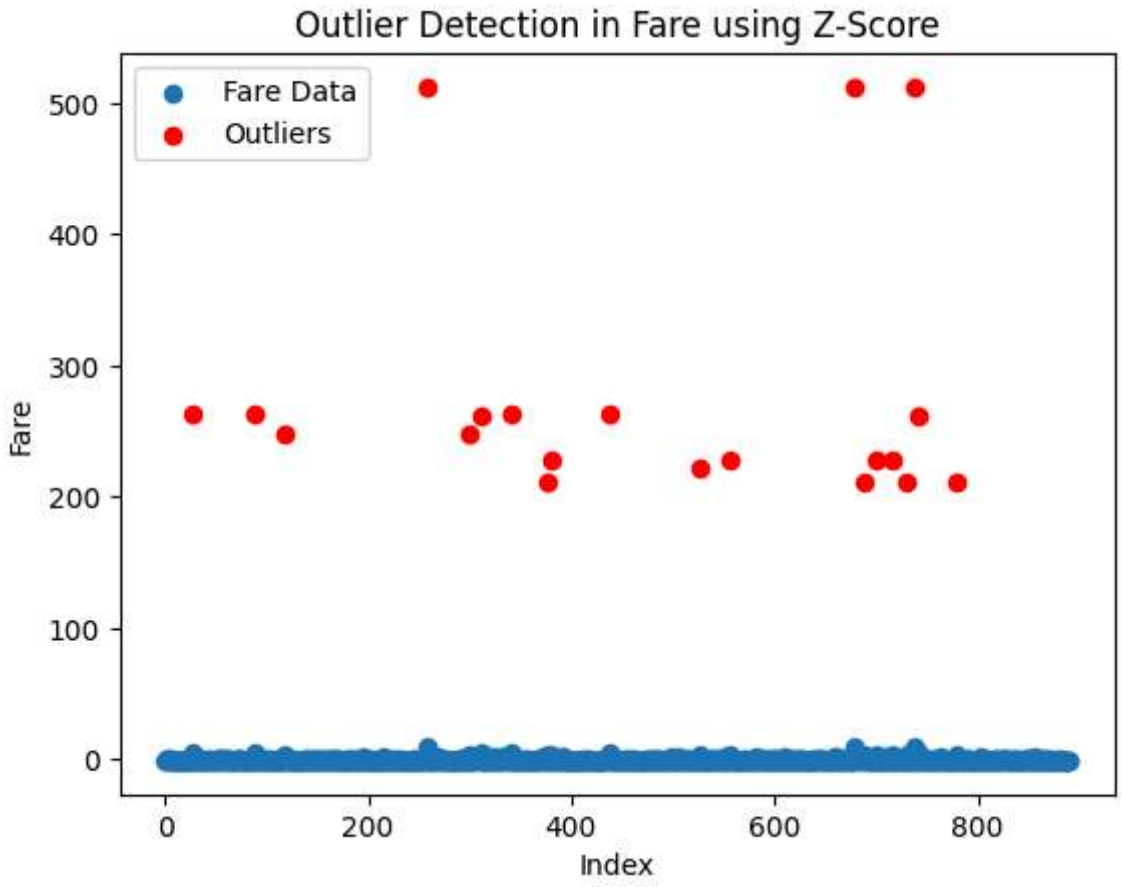
```
Passenger 1: Survived
Passenger 2: Survived
Passenger 3: Survived
```

In [25]:
```
import pandas as pd
import numpy as np
titanic_df = sns.load_dataset('titanic')
# Outlier detection - 'Age'
mean_age = np.mean(titanic_df['age']) # calculates the mean
std_dev_age = np.std(titanic_df['age']) # calculates the standard deviation
Z_scores_age = (titanic_df['age'] - mean_age) / std_dev_age # computes the Z-scores
outliers_age = titanic_df['age'][np.abs(Z_scores_age) > 3] # finds all the data poi
print("Outliers in 'Age' using Z-score: \n", outliers_age)

# Outlier detection - 'Fare'
mean_fare = np.mean(titanic_df['fare']) # calculates the mean
std_dev_fare = np.std(titanic_df['fare']) # calculates the standard deviation
Z_scores_fare = (titanic_df['fare'] - mean_fare) / std_dev_fare # computes the Z-sc
outliers_fare = titanic_df['fare'][np.abs(Z_scores_fare) > 3] # finds all the data
print("\nOutliers in 'Fare' using Z-score: \n", outliers_fare)
```

```
Outliers in 'Age' using Z-score:
 630    80.0
851    74.0
Name: age, dtype: float64

Outliers in 'Fare' using Z-score:
 27     263.0000
88     263.0000
118    247.5208
258    512.3292
299    247.5208
311    262.3750
341    263.0000
377    211.5000
380    227.5250
438    263.0000
527    221.7792
557    227.5250
679    512.3292
689    211.3375
700    227.5250
716    227.5250
730    211.3375
737    512.3292
742    262.3750
779    211.3375
Name: fare, dtype: float64
```

In [30]:
```
import matplotlib.pyplot as plt

plt.figure()
plt.scatter(titanic_df.index, titanic_df['age'], label='Age Data')
plt.scatter(outliers_age.index, outliers_age, color='red', label='Outliers')
plt.title("Outlier Detection in Age using Z-Score")
plt.xlabel("Index")
plt.ylabel("Age")
plt.legend()
plt.show()
plt.figure()
plt.scatter(titanic_df.index, titanic_df['fare'], label='Fare Data')
plt.scatter(outliers_fare.index, outliers_fare, color='red', label='Outliers')
plt.title("Outlier Detection in Fare using Z-Score")
plt.xlabel("Index")
plt.ylabel("Fare")
plt.legend()
plt.show()
```

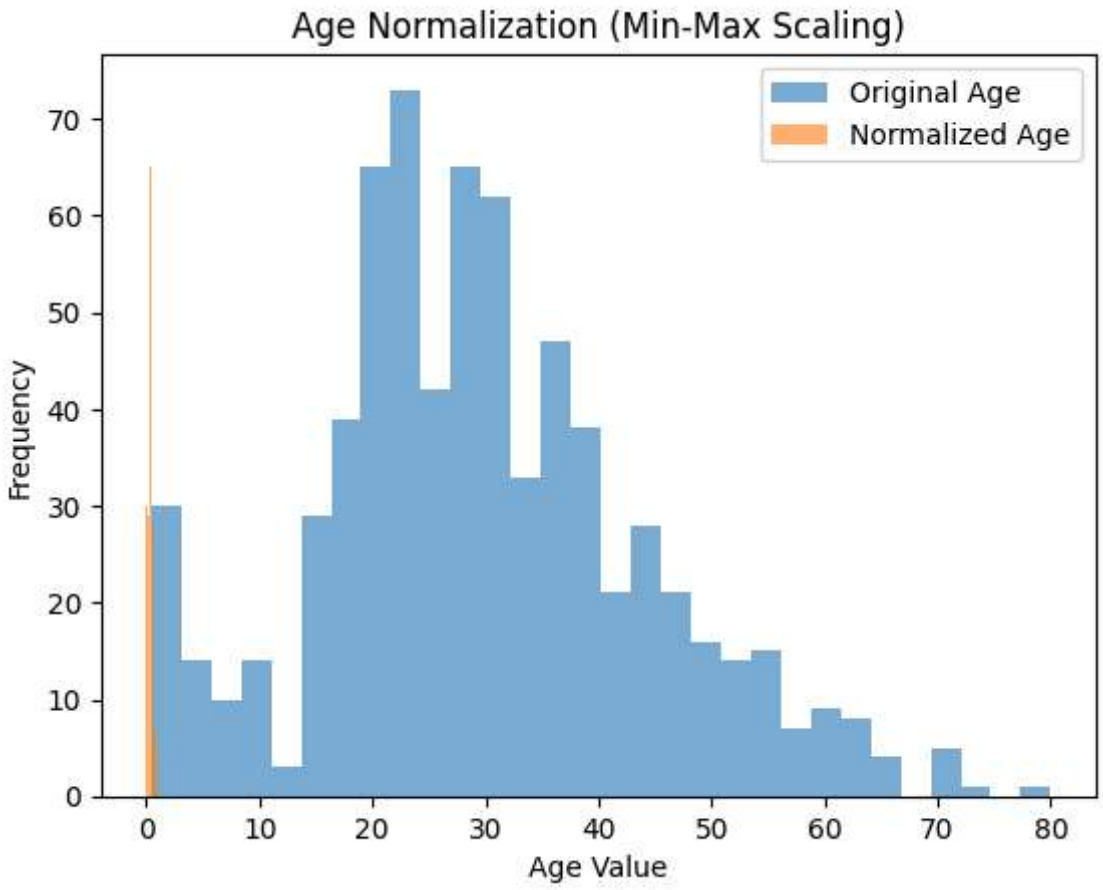

Outlier Detection in Age using Z-Score
Age Data
Outliers
Age
Index
0
10
20
30
40
50
60
70
80
0
200
400
600
800


Outlier Detection in Fare using Z-Score
Fare Data
Outliers
Fare
Index
0
100
200
300
400
500
0
200
400
600
800

In [26]:
```
# Import necessary libraries
import seaborn as sns
import pandas as pd

# Load the Titanic Dataset


# Normalize 'age'
titanic_df['age'] = (titanic_df['age'] - titanic_df['age'].min()) / (titanic_df['ag

# Display the normalized ages
print(titanic_df['age'])
```

```
0      0.271174
1      0.472229
2      0.321438
3      0.434531
4      0.434531
         ...
886    0.334004
887    0.233476
888         NaN
889    0.321438
890    0.396833
Name: age, Length: 891, dtype: float64
```

In [31]:
```
import matplotlib.pyplot as plt

# Re-load original age for comparison
original_age = sns.load_dataset('titanic')['age']

plt.figure()
plt.hist(original_age.dropna(), bins=30, alpha=0.6, label='Original Age')
plt.hist(titanic_df['age'].dropna(), bins=30, alpha=0.6, label='Normalized Age')
plt.title("Age Normalization (Min-Max Scaling)")
plt.xlabel("Age Value")
plt.ylabel("Frequency")
plt.legend()
plt.show()
```

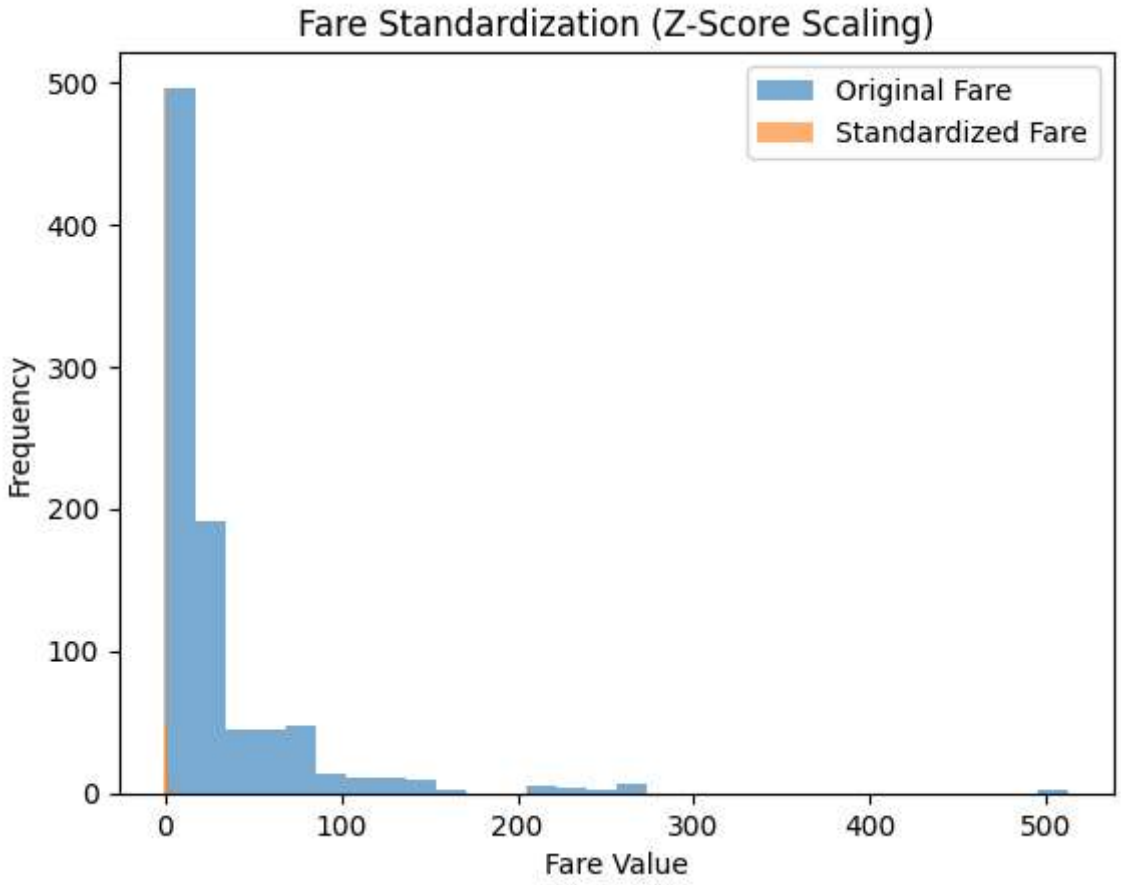


In [27]:
```
# Standardize 'fare'
titanic_df['fare'] = (titanic_df['fare'] - titanic_df['fare'].mean()) / titanic_df[

# Display the standardized fares
print(titanic_df['fare'])
```

```
0     -0.502163
1      0.786404
2     -0.488580
3      0.420494
4     -0.486064
         ...
886   -0.386454
887   -0.044356
888   -0.176164
889   -0.044356
890   -0.492101
Name: fare, Length: 891, dtype: float64
```

In [32]:
```
import matplotlib.pyplot as plt

# Re-load original fare for comparison
original_fare = sns.load_dataset('titanic')['fare']

plt.figure()
plt.hist(original_fare, bins=30, alpha=0.6, label='Original Fare')
plt.hist(titanic_df['fare'], bins=30, alpha=0.6, label='Standardized Fare')
plt.title("Fare Standardization (Z-Score Scaling)")
```

```
plt.xlabel("Fare Value")
plt.ylabel("Frequency")
plt.legend()
plt.show()
```



In [ ]: