Visualize_Distribution

September 6, 2024

1 Đọc vào dữ liệu sau khi qua module 1

[ ]:
```
import pandas as pd

# Đọc dữ liệu từ CSV
data_dir = '../../../data/processed/'
train_df = pd.read_csv(data_dir + 'train_features.csv').
 ↪drop(columns=['image_path'])
valid_df = pd.read_csv(data_dir + 'valid_features.csv').
 ↪drop(columns=['image_path'])
test_df = pd.read_csv(data_dir + 'test_features.csv').
 ↪drop(columns=['image_path'])

# Xem cấu trúc của DataFrame
print(train_df.head())
```

```
   label     dim_0     dim_1     dim_2     dim_3     dim_4     dim_5  \
0      2  1.518592 -2.122206  1.063359  1.083591 -3.078244  1.832143
1      2 -3.997844 -2.013066  5.606269 -2.221863  0.908517  0.995050
2      2 -4.144322 -3.335181  7.014940 -2.851661 -1.254374  4.090703
3      2 -1.838009 -3.809826  5.647745 -4.356337  0.269498  3.153944
4      2 -6.741322 -0.116311  5.371378 -3.934162  2.132378  2.241241

      dim_6     dim_7     dim_8  …    dim_29    dim_30    dim_31    dim_32  \
0 -3.617894  1.835063  0.739837  … -0.450482  0.230301  0.322319 -0.361697
1 -4.953275  1.075495  3.199515  …  2.695855 -4.411282 -1.748895  5.605376
2 -3.888599 -2.345787  6.091836  …  6.469357 -5.430499 -2.013651  6.954945
3 -2.280085  0.251856  1.419379  …  4.250416 -5.249067 -1.796968  6.298747
4 -5.419630  2.448615  2.177936  …  2.325113 -4.234541  1.869896  1.142828

     dim_33    dim_34    dim_35    dim_36    dim_37    dim_38
0 -1.096648  1.402599  0.442939 -3.928780  2.343227  1.505659
1 -6.375985 -0.828269  6.179725 -4.965612  0.590781  3.893331
2 -3.220028 -1.721707  5.068684 -4.063029 -1.196603  5.176606
3 -3.189471 -3.151461  5.499859 -6.258601 -0.574315  6.065775
4 -6.010725  1.315534  3.703765 -3.615420  0.399788  2.430184

[5 rows x 40 columns]
```

2 Trực quan hóa phân phối của dữ liệu theo từng nhãn

2.1 Code chức năng

[ ]:
```python
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Sử dụng box plot
def visualize_distribution(features, labels):
    # Create a figure with 4 subplots
    fig, axs = plt.subplots(2, 2, figsize=(10, 10))
    fig.suptitle("Data Distribution Visualizations")

    # Plot for all data
    sns.boxplot(data=features, ax=axs[1, 1])
    axs[1, 1].set_title("All Data")
    axs[1, 1].set_xlabel("Features")
    axs[1, 1].set_ylabel("Values")

    # Plot for each label
    for i, label in enumerate([0, 1, 2]):
        label_data = features[labels == label]
        sns.boxplot(data=label_data, ax=axs[i // 2, i % 2])
        axs[i // 2, i % 2].set_title(f"Label {label}")
        axs[i // 2, i % 2].set_xlabel("Features")
        axs[i // 2, i % 2].set_ylabel("Values")

    plt.tight_layout()
    plt.show()


# Sử dụng histogram
def visualize_distribution_histogram(features, labels):
    # Create 4 separate figures
    num_features = features.shape[1]
    for i in range(4):
        if num_features == 3:
            plt.figure(figsize=(10, 10))
        else:
            plt.figure(figsize=(15, 15))

        if i == 3:
            title = "All Data"
            data_to_plot = features
        else:
            title = f"Class {i}"
```

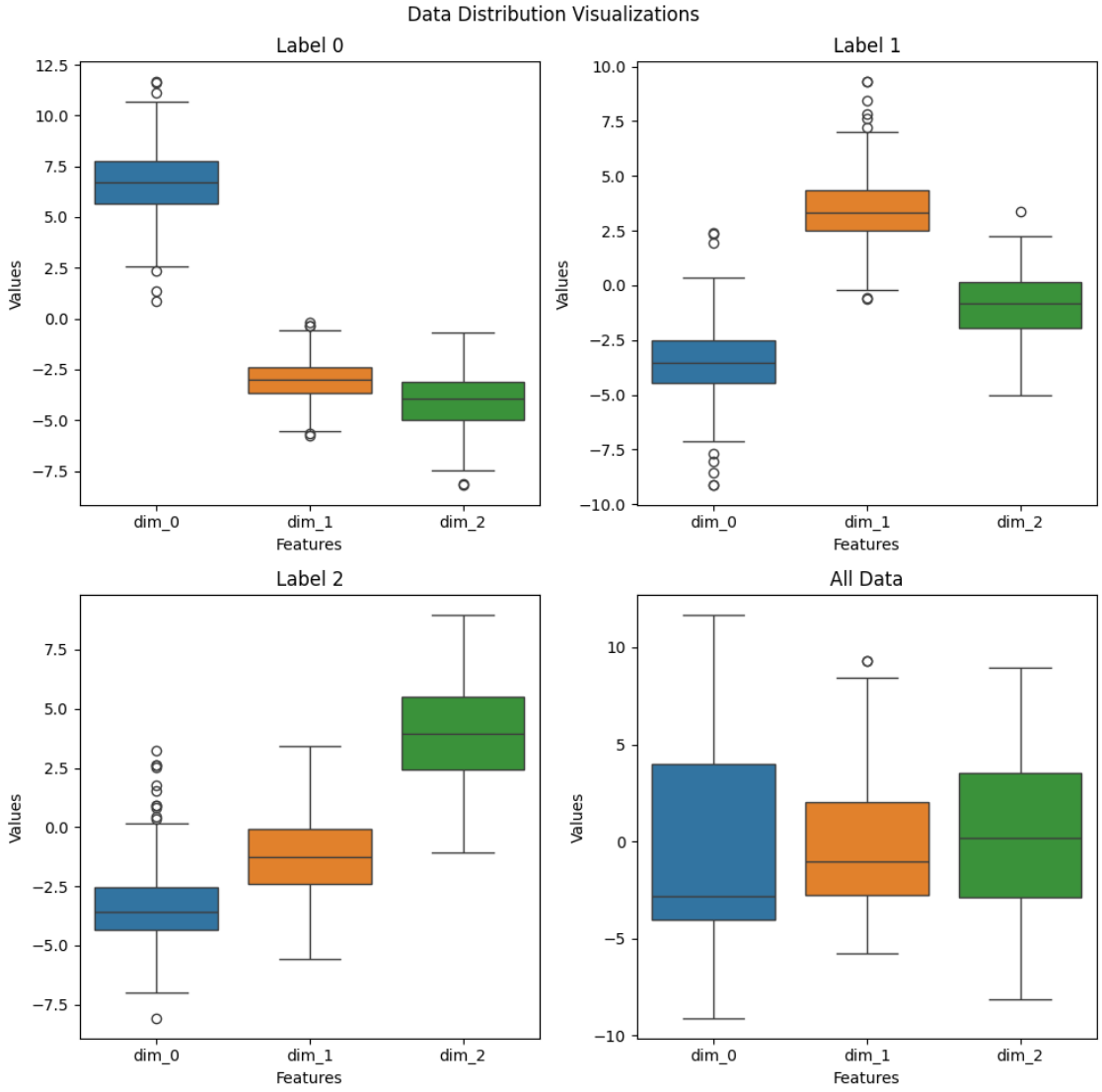
```
            data_to_plot = features[labels == i]

        plt.suptitle(f"Data Distribution: {title}")

        sns.histplot(data_to_plot, kde=True, binwidth=0.75)
        # if num_features == 3:
        #     sns.histplot(data_to_plot, kde=True, binwidth=1)
        # else:
        #     sns.histplot(data_to_plot, kde=True, binwidth=2)
        plt.xlabel("Value")
        plt.ylabel("Count")

        plt.tight_layout()
        plt.show()
```

2.2 Với chỉ 3 chiều đầu tiên

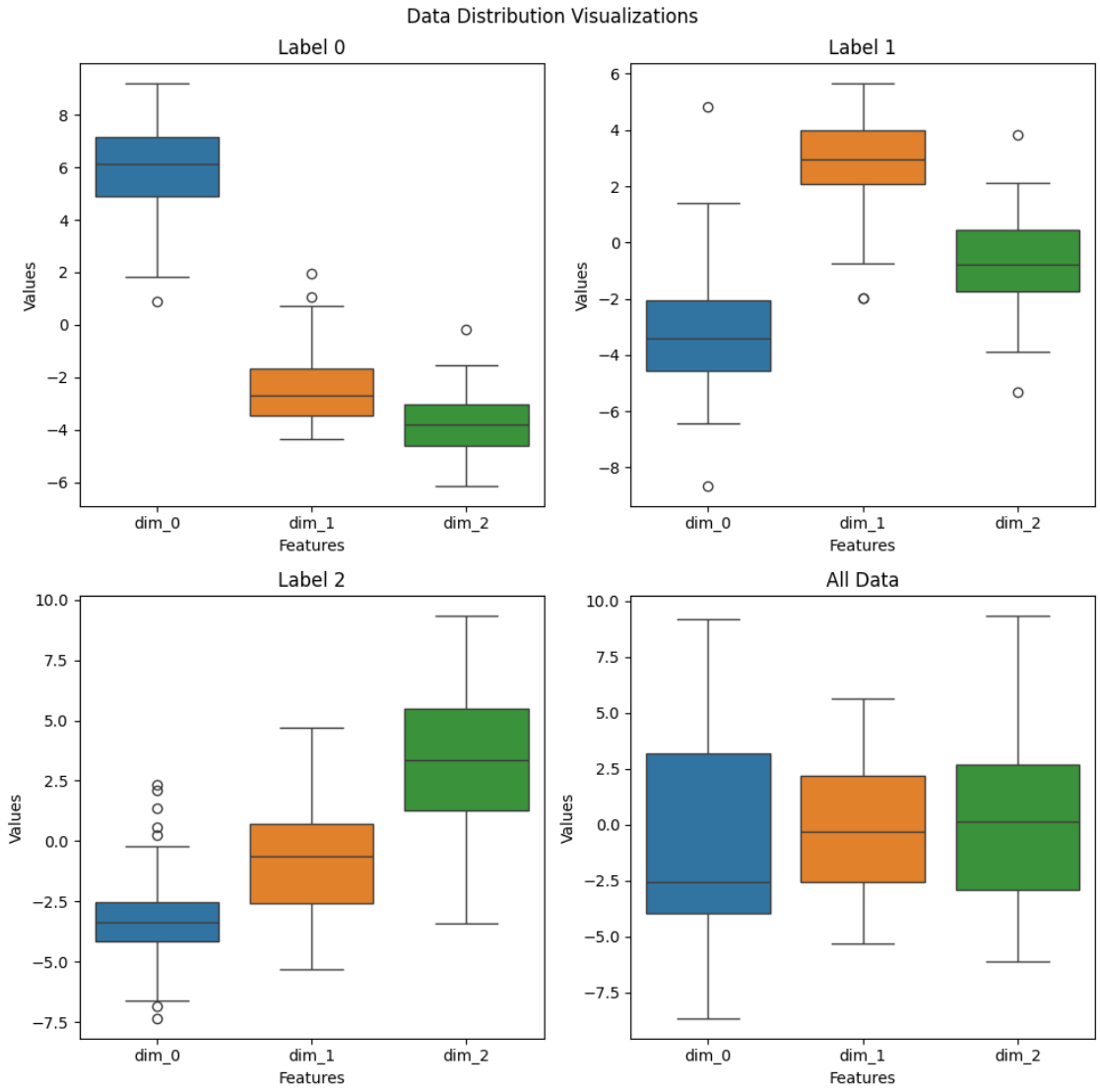
[ ]:
```
# Separate features and labels
train_features_3 = train_df.iloc[:, 1:4]
train_labels_3 = train_df.iloc[:, 0]
valid_features_3 = valid_df.iloc[:, 1:4]
valid_labels_3 = valid_df.iloc[:, 0]
test_features_3 = test_df.iloc[:, 1:4]
test_labels_3 = test_df.iloc[:, 0]
```

[ ]:
```
print("Train visualize")
visualize_distribution(train_features_3, train_labels_3)
print("Validation visualize")
visualize_distribution(valid_features_3, valid_labels_3)
print("Test visualize")
visualize_distribution(test_features_3, test_labels_3)
```

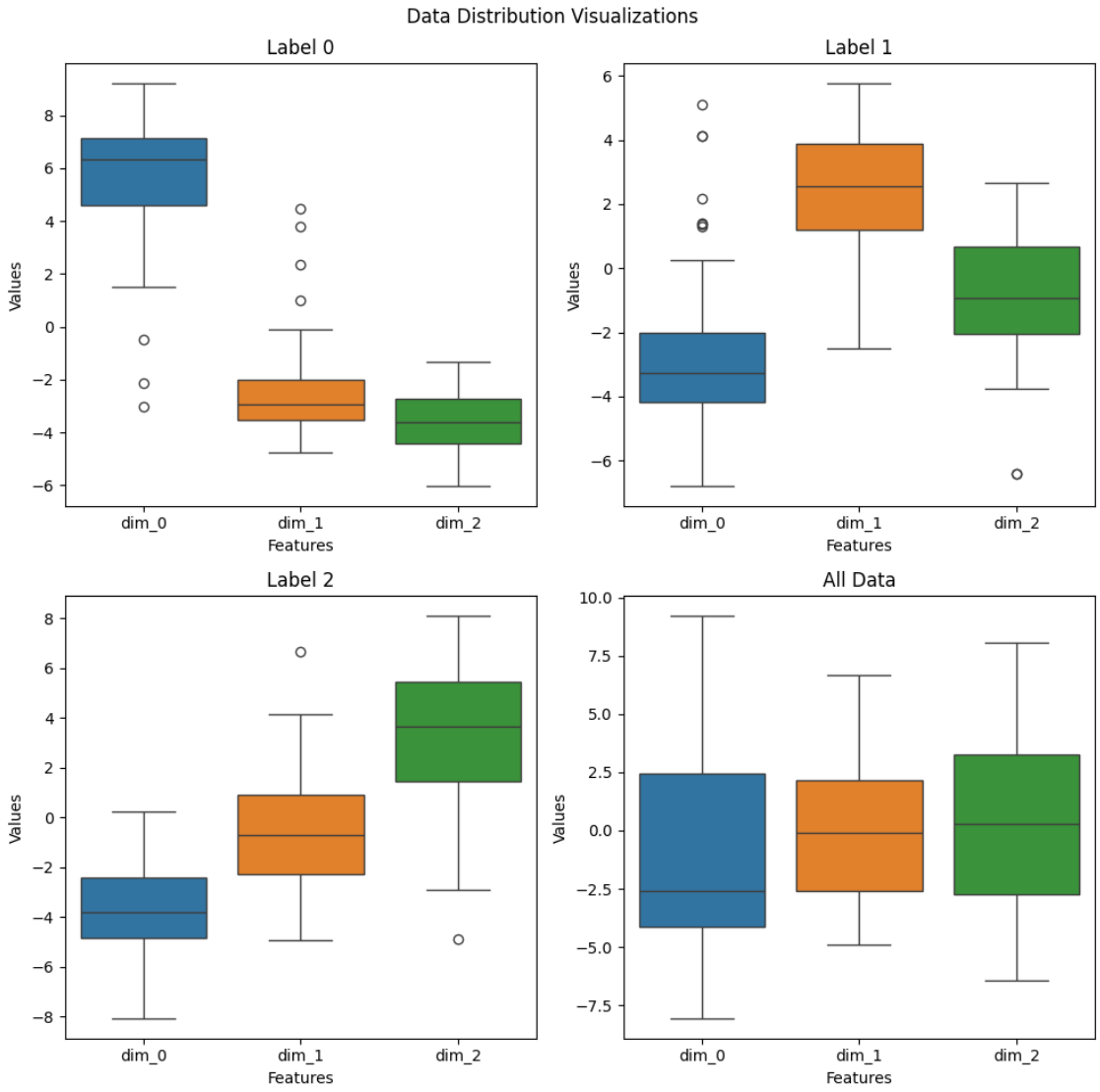
```
Train visualize
```



`Validation visualize`

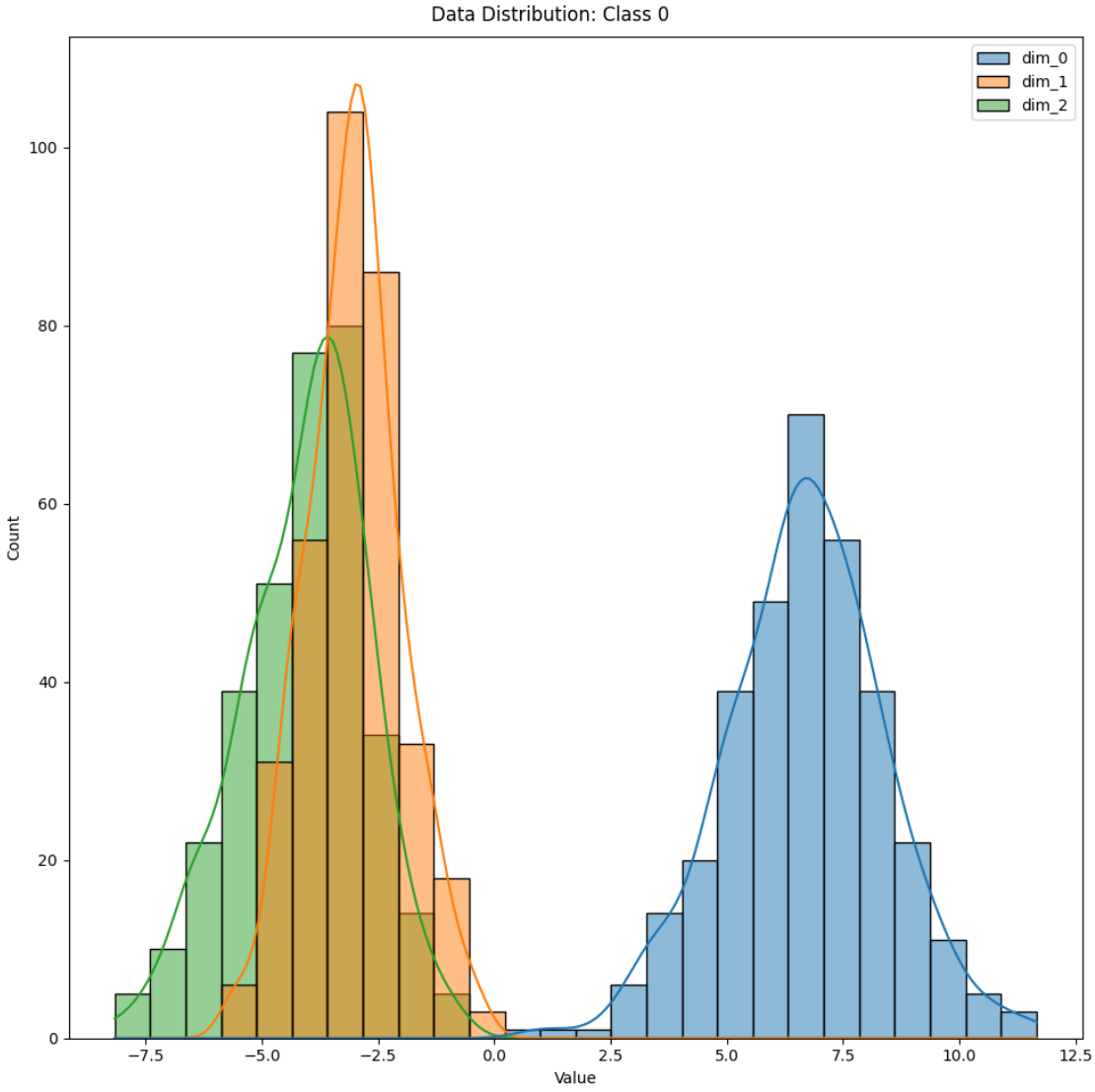


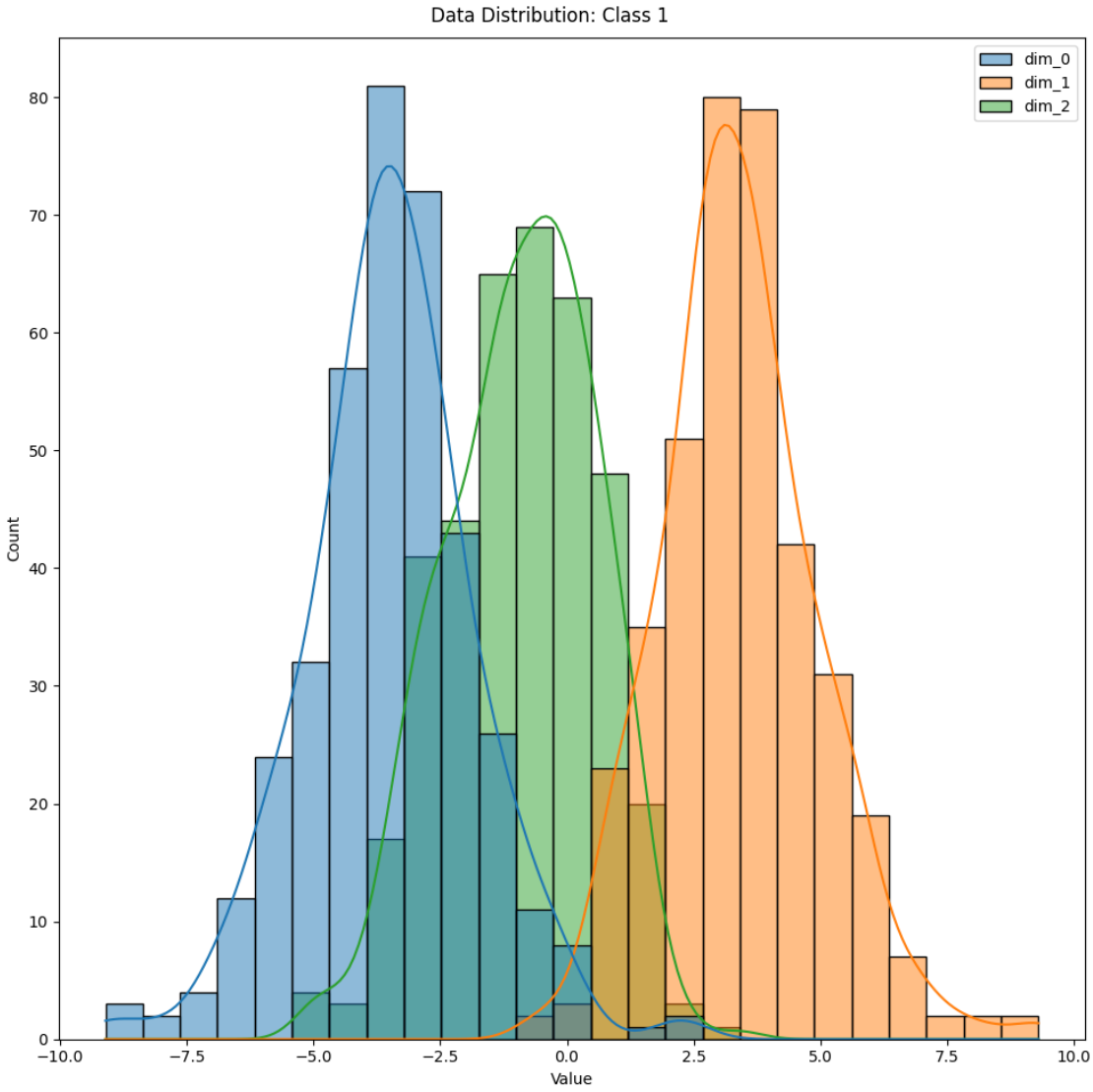

```
Test visualize
```

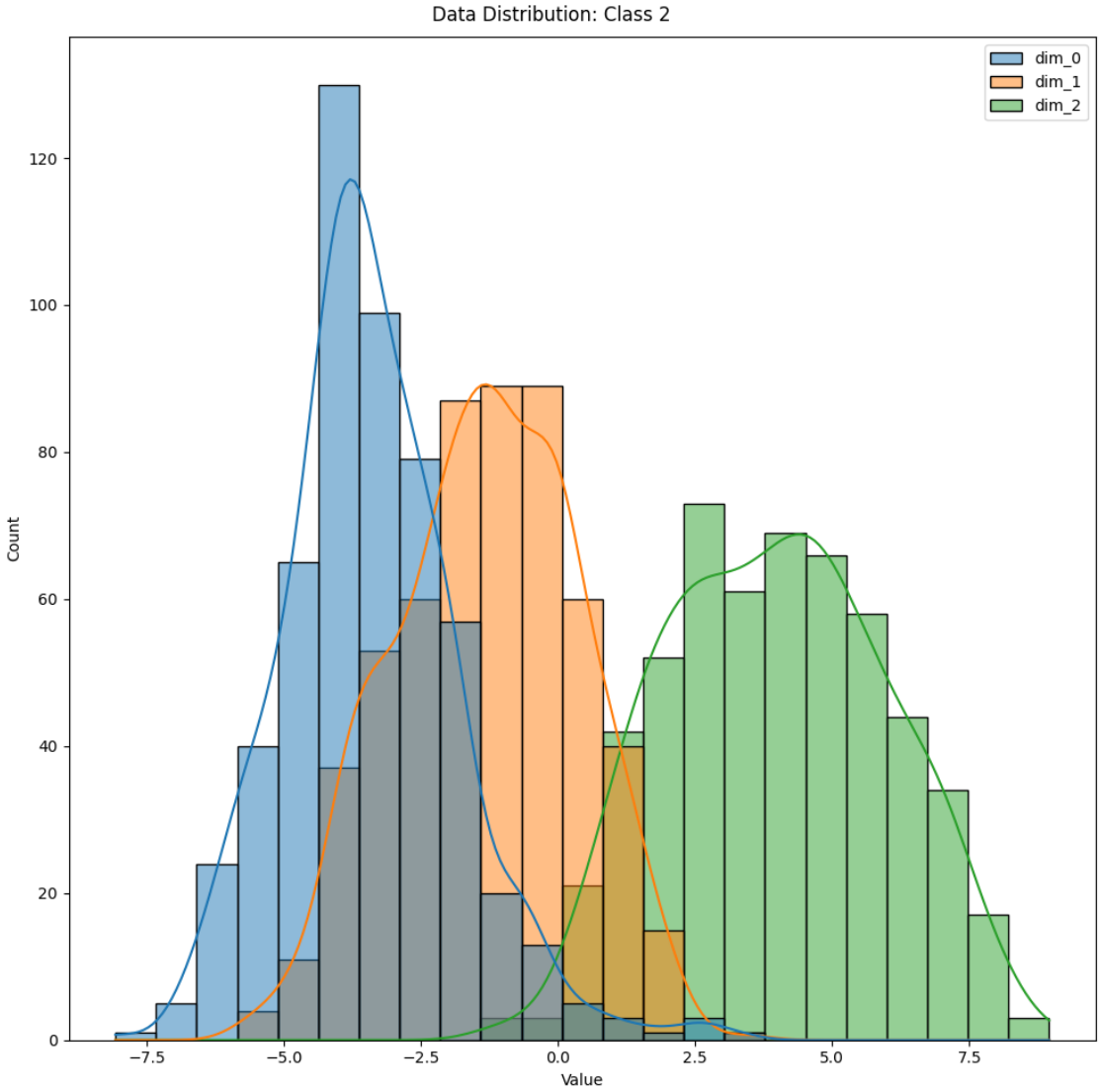


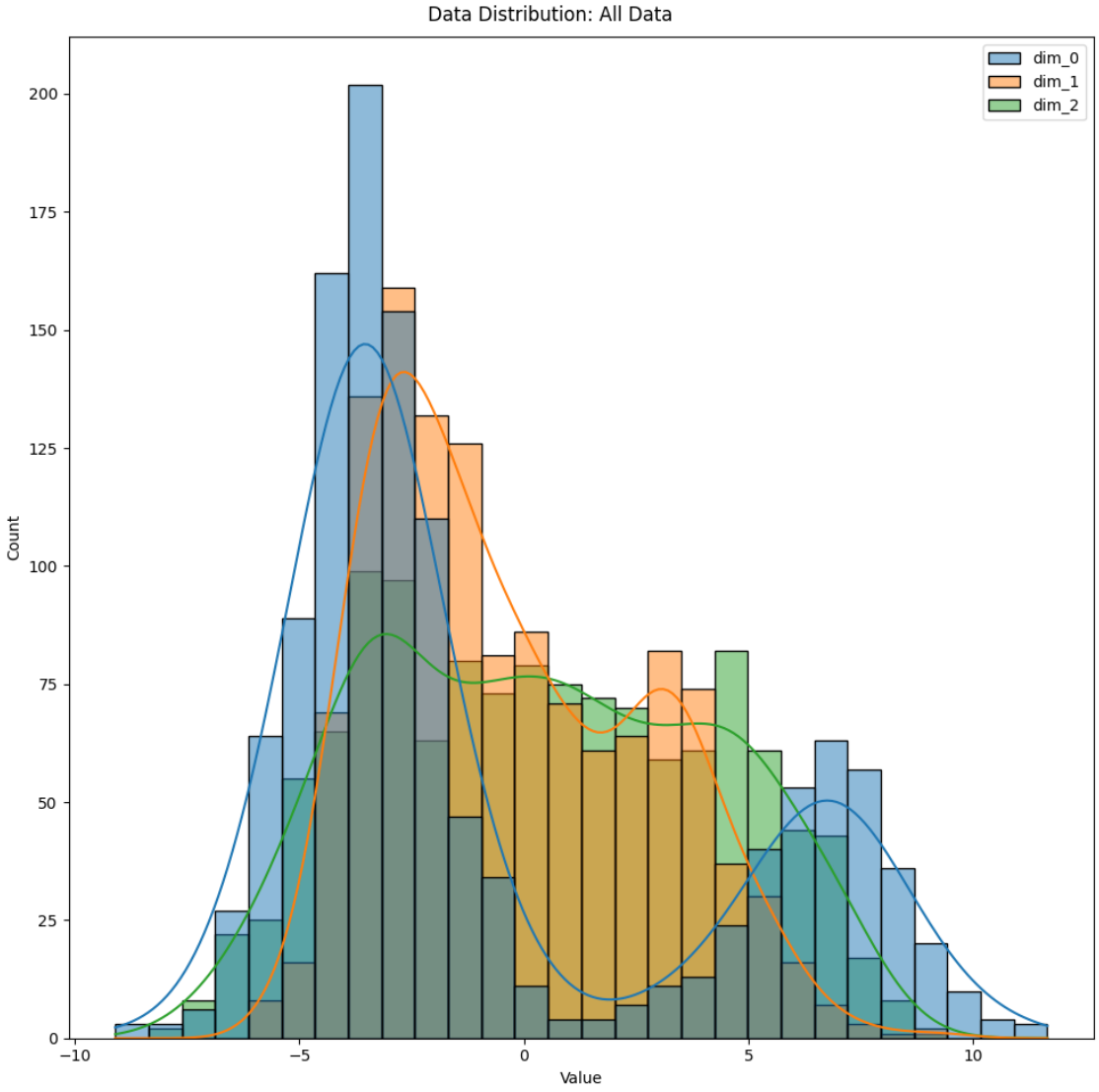
[ ]:
```
print("Train visualize")
visualize_distribution_histogram(train_features_3, train_labels_3)
print("Validation visualize")
visualize_distribution_histogram(valid_features_3, valid_labels_3)
print("Test visualize")
visualize_distribution_histogram(test_features_3, test_labels_3)
```

```
Train visualize
```

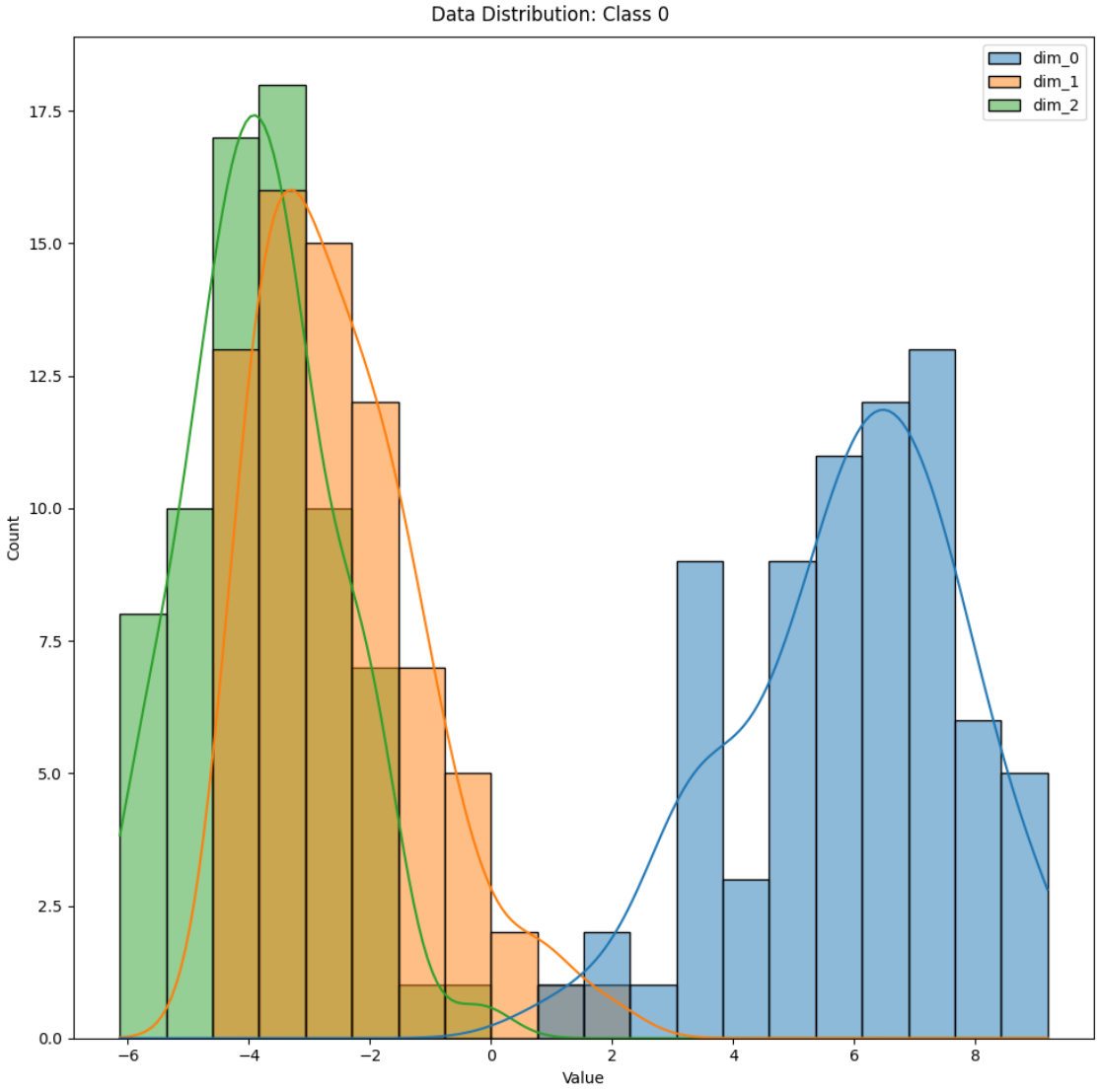

Data Distribution: Class 0
dim_0
dim_1
dim_2
Count
Value
0
20
40
60
80
100
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0
12.5

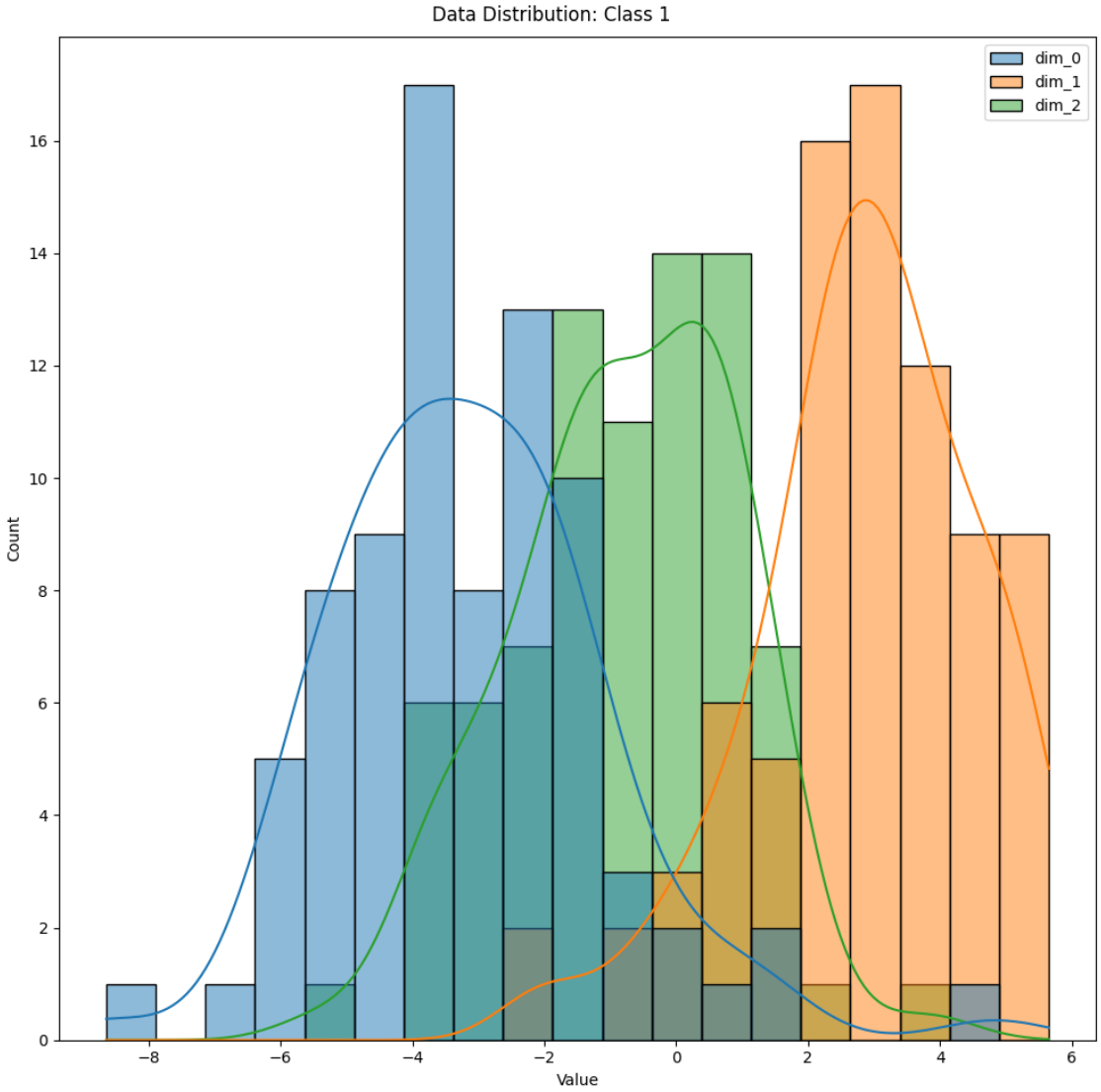

Data Distribution: Class 1
dim_0
dim_1
dim_2
Count
Value
0
10
20
30
40
50
60
70
80
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0

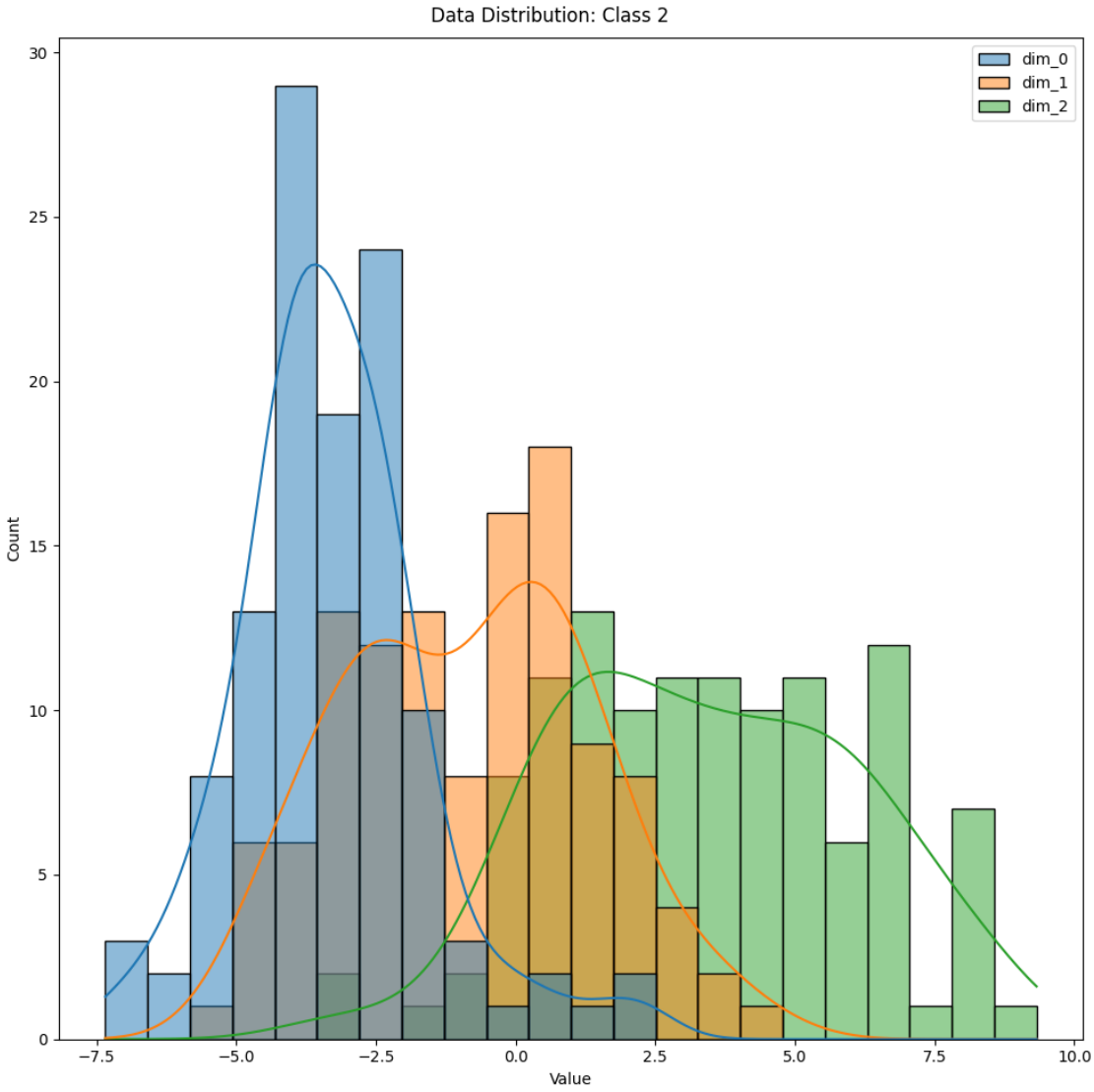

Data Distribution: Class 2
dim_0
dim_1
dim_2
Count
Value
0
20
40
60
80
100
120
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5

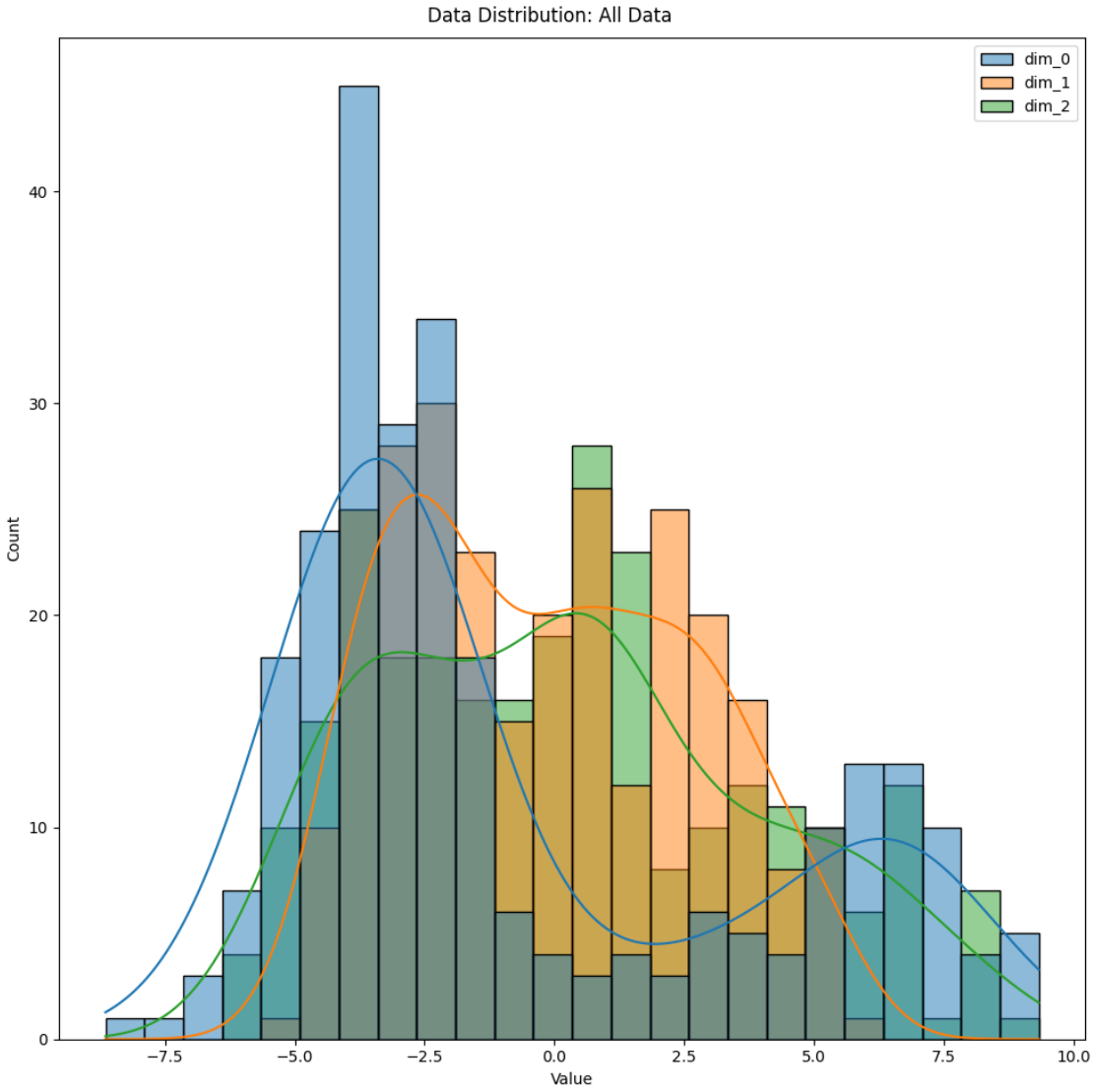


Validation visualize

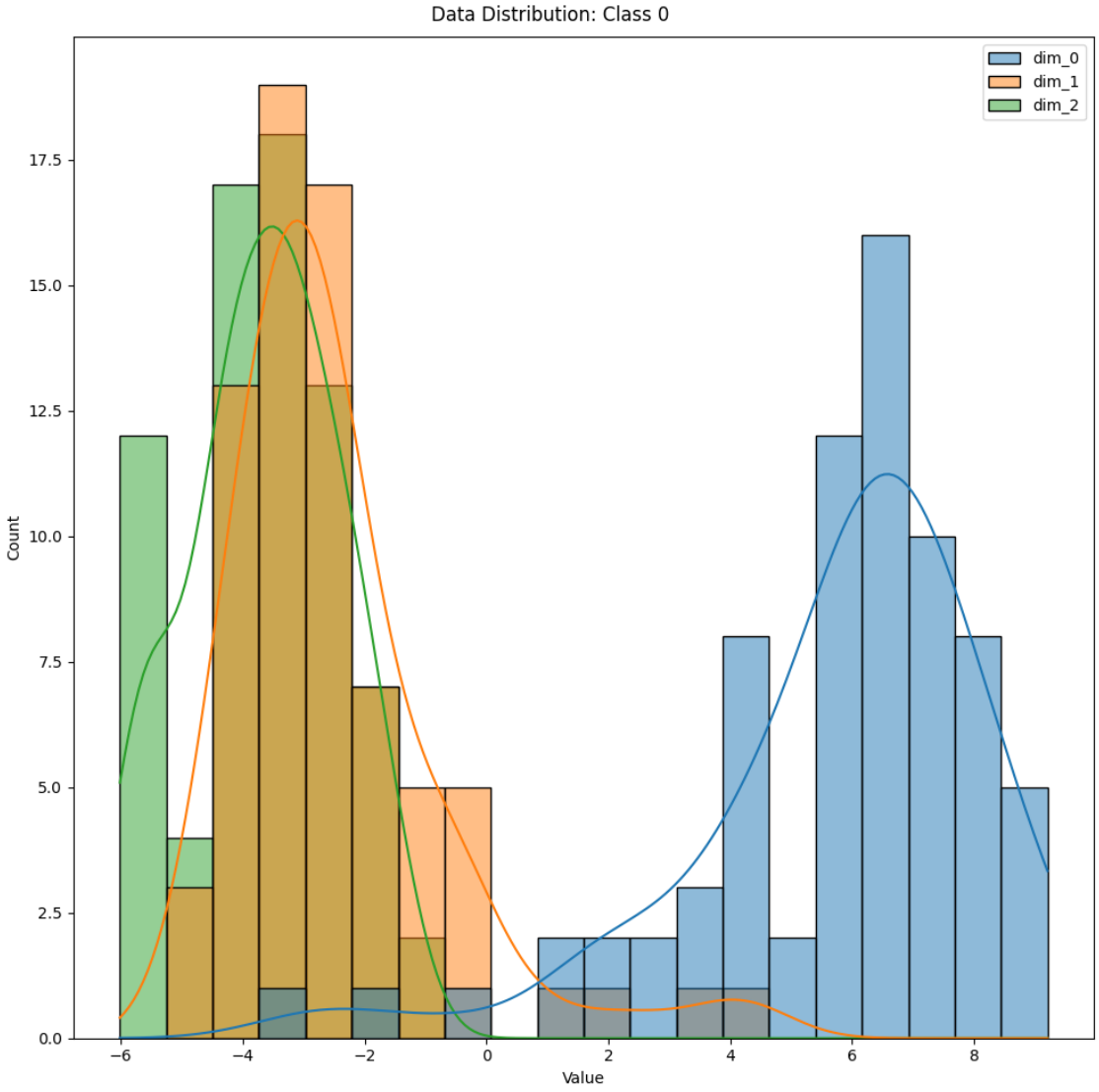

Data Distribution: Class 0
dim_0
dim_1
dim_2
Count
Value
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
−6
−4
−2
0
2
4
6
8

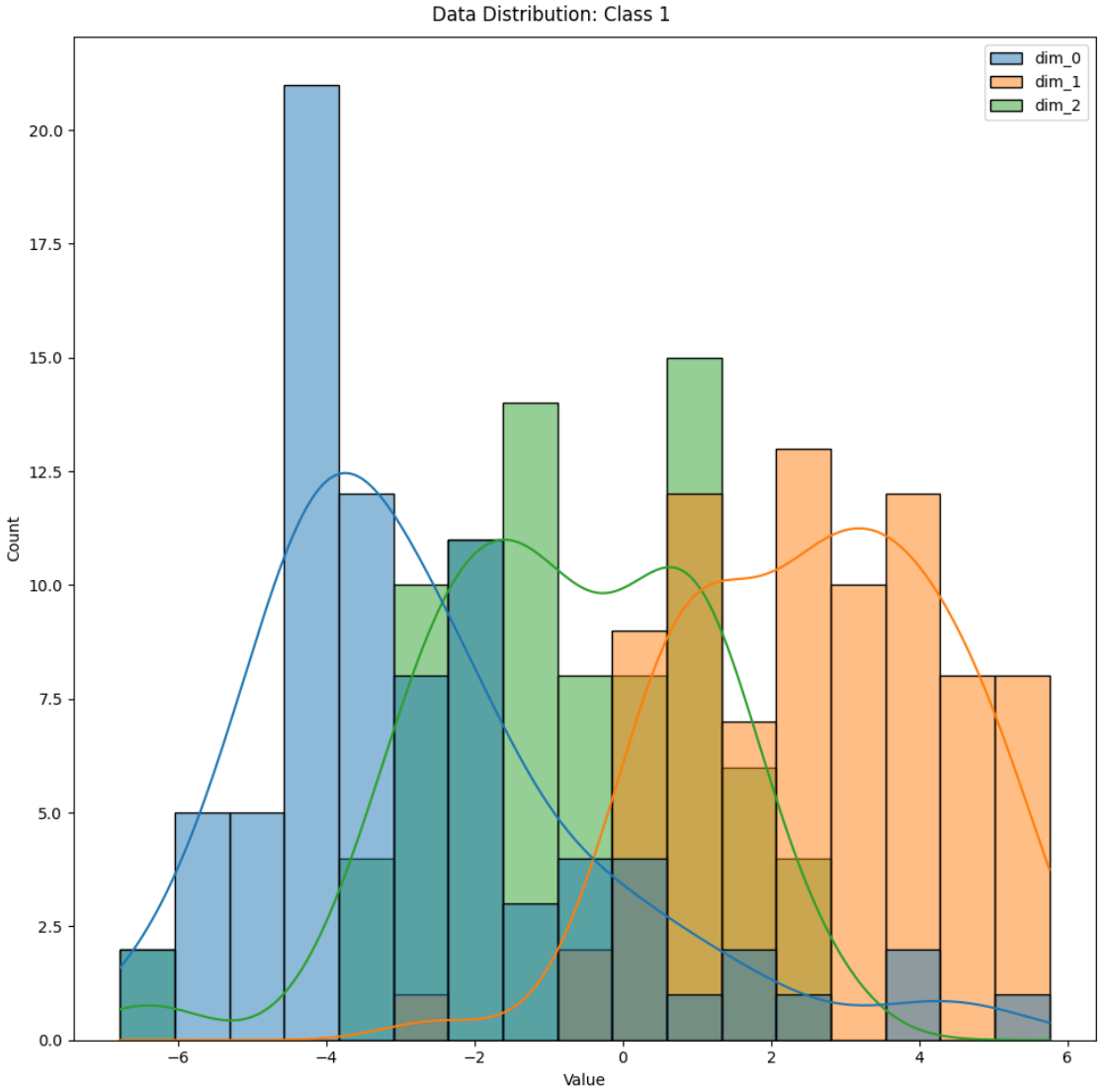

Data Distribution: Class 1
dim_0
dim_1
dim_2
Count
Value
0
2
4
6
8
10
12
14
16
−8
−6
−4
−2
0
2
4
6

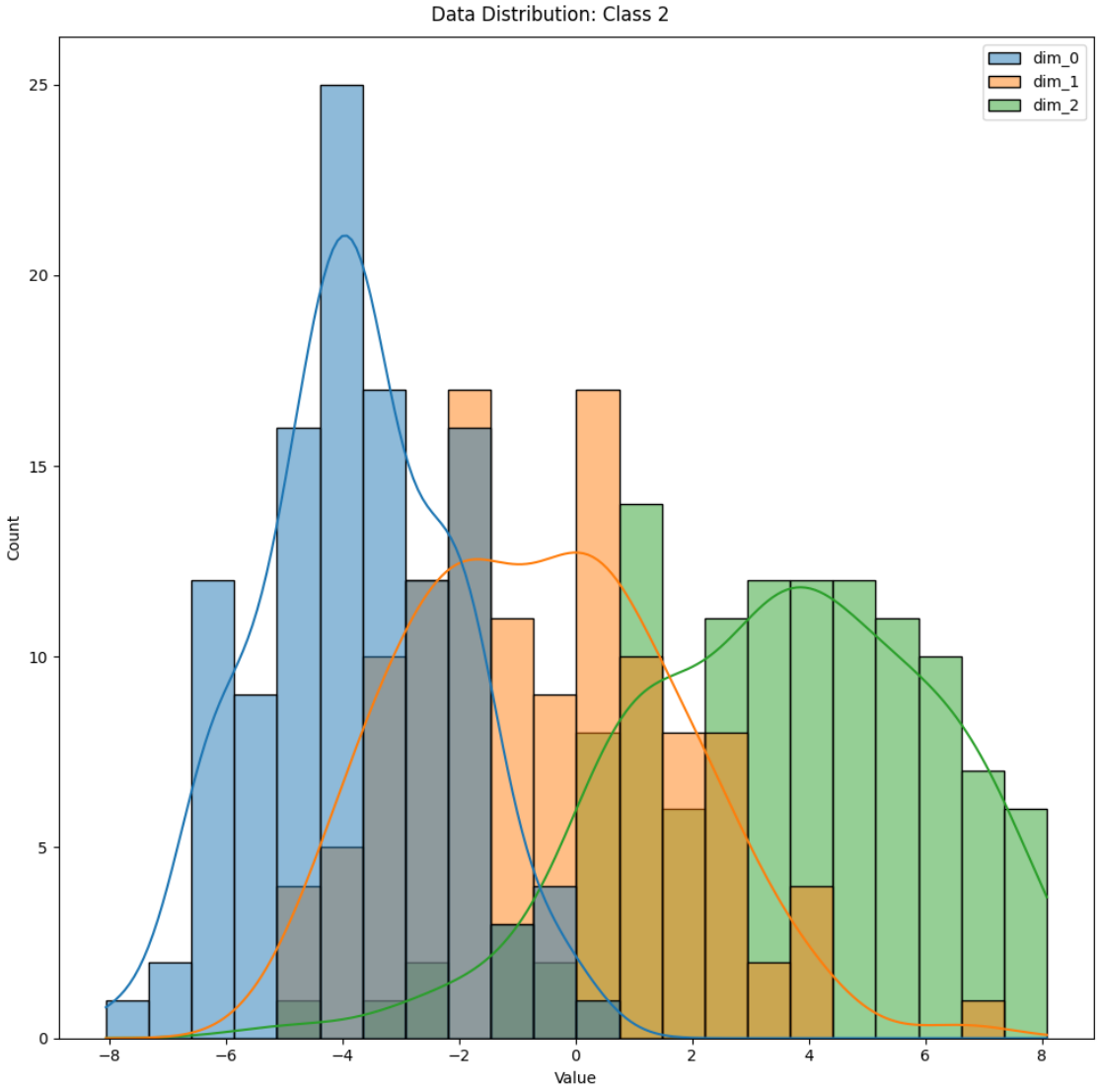

Data Distribution: Class 2
dim_0
dim_1
dim_2
Count
Value
0
5
10
15
20
25
30
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0

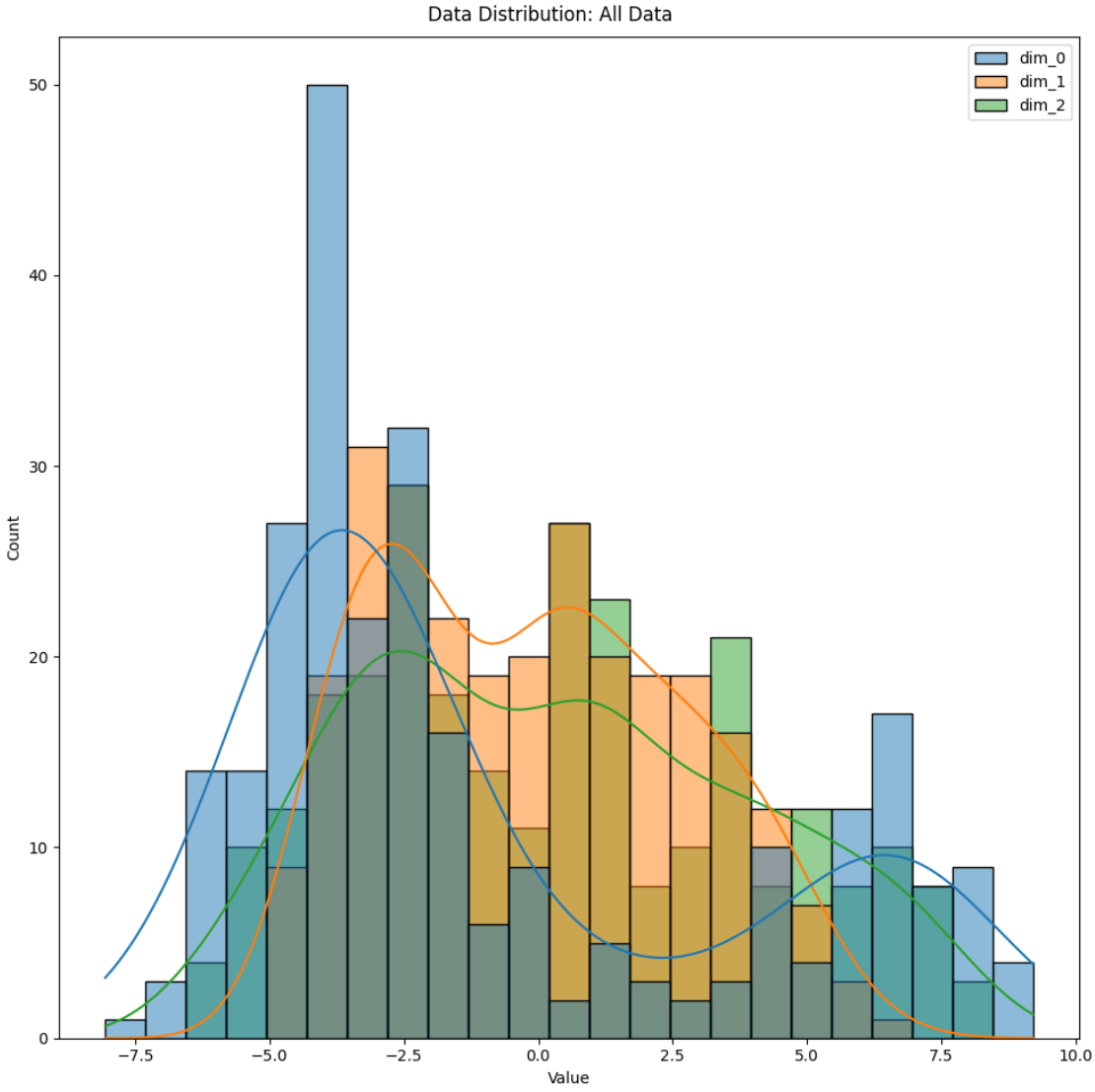



```
Test visualize
```


Data Distribution: Class 0
dim_0
dim_1
dim_2
Count
Value
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
−6
−4
−2
0
2
4
6
8


Data Distribution: Class 1
dim_0
dim_1
dim_2
Count
Value
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
−6
−4
−2
0
2
4
6


Data Distribution: Class 2
dim_0
dim_1
dim_2
Count
Value
0
5
10
15
20
25
−8
−6
−4
−2
0
2
4
6
8



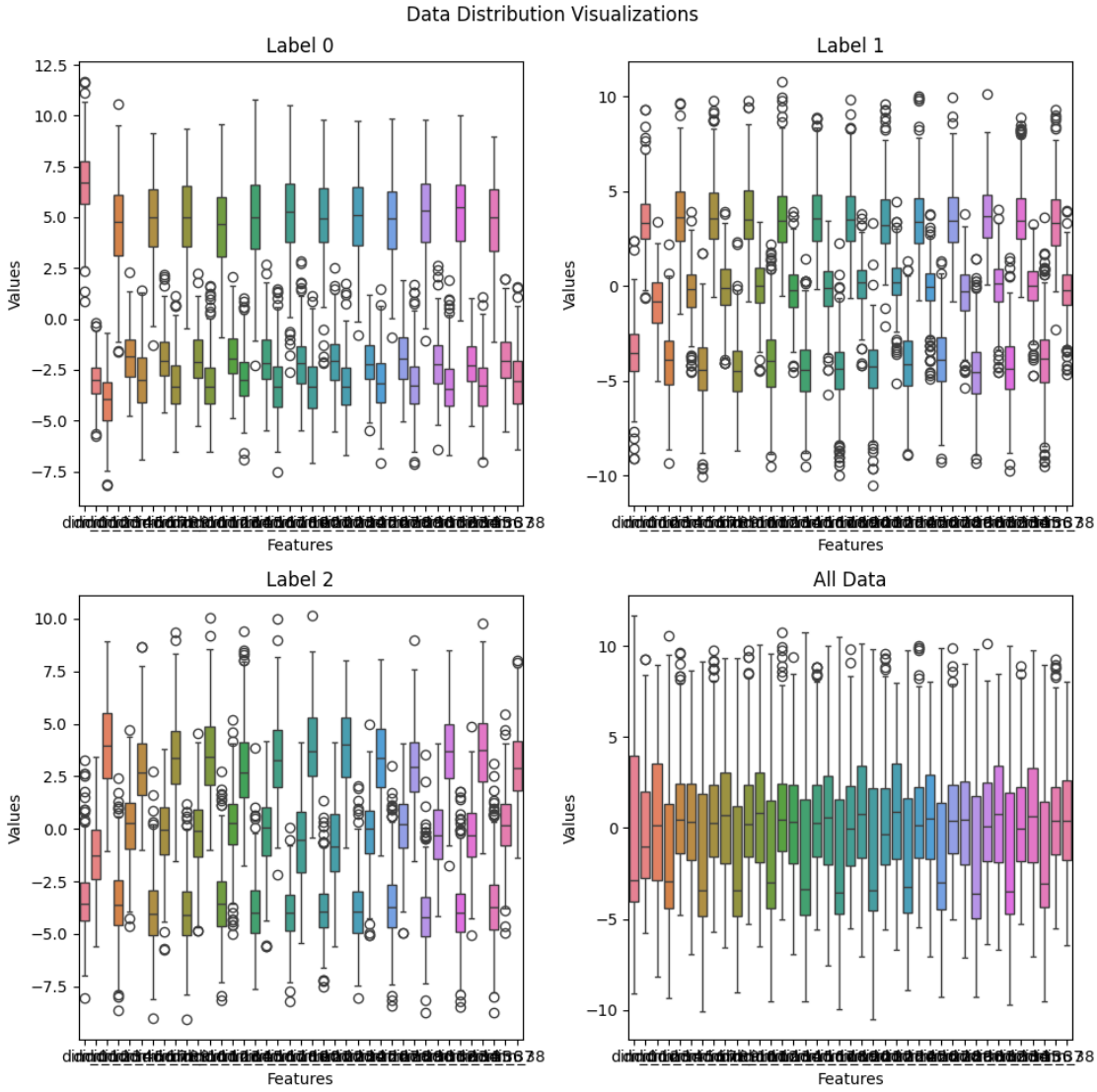
2.3 Với tất cả 39 chiều

[ ]:
```
# Separate features and labels
train_features_39 = train_df.iloc[:, 1:40]
train_labels_39 = train_df.iloc[:, 0]
valid_features_39 = valid_df.iloc[:, 1:40]
valid_labels_39 = valid_df.iloc[:, 0]
test_features_39 = test_df.iloc[:, 1:40]
test_labels_39 = test_df.iloc[:, 0]
```

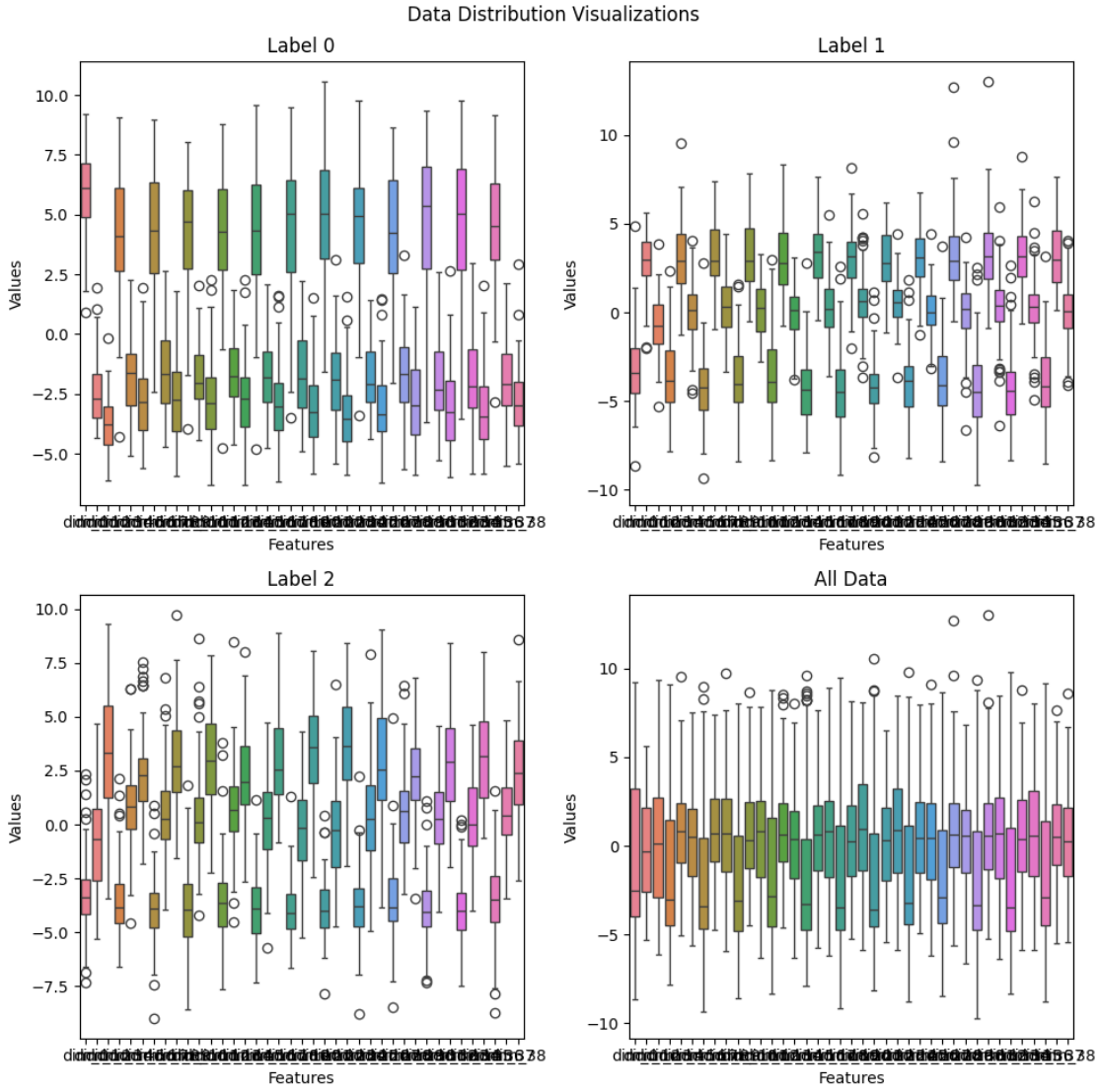
[ ]:
```
print("Train visualize")
visualize_distribution(train_features_39, train_labels_39)
print("Validation visualize")
```

```
visualize_distribution(valid_features_39, valid_labels_39)
print("Test visualize")
visualize_distribution(test_features_39, test_labels_39)
```

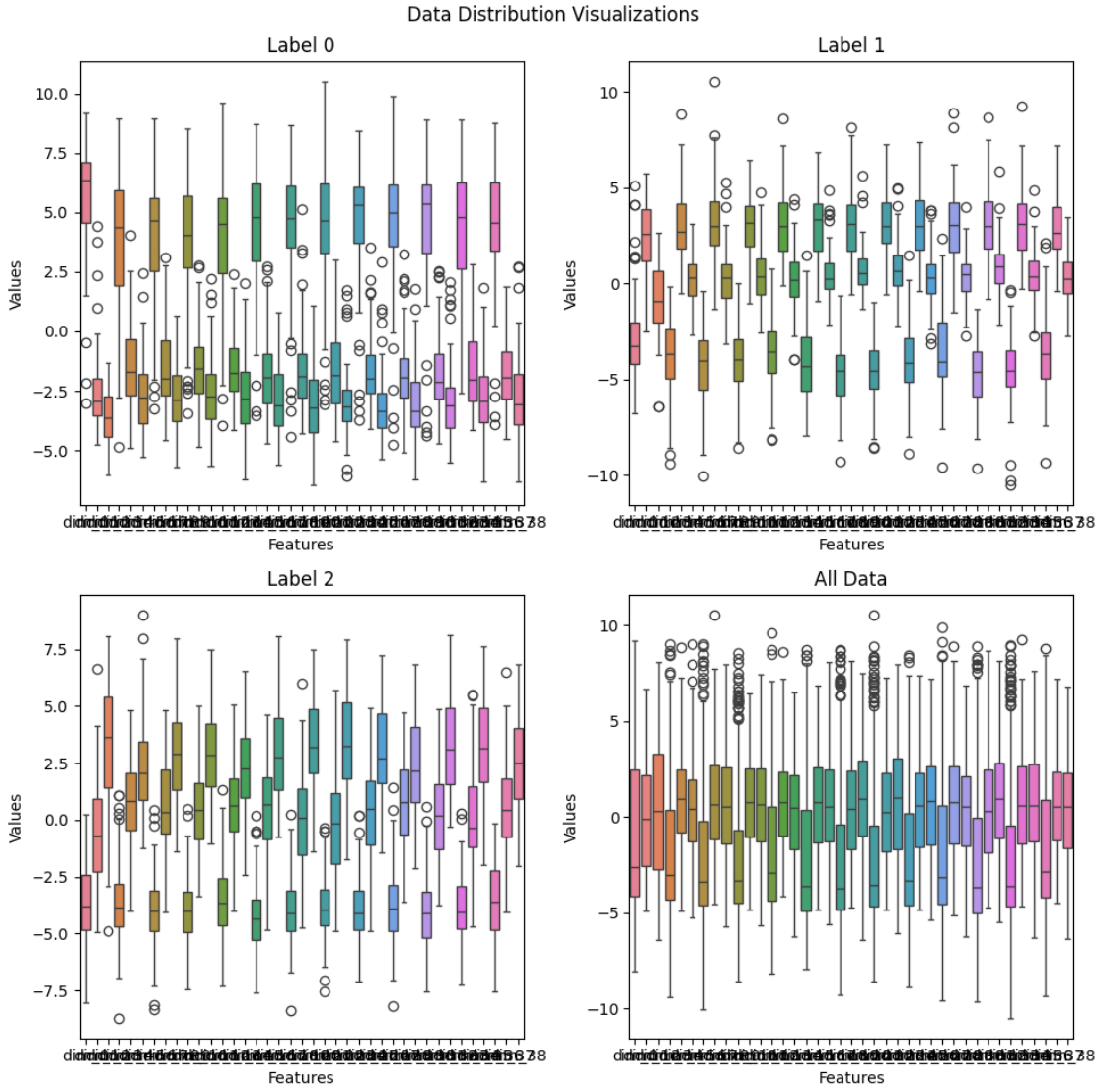
```
Train visualize
```




```
Validation visualize
```

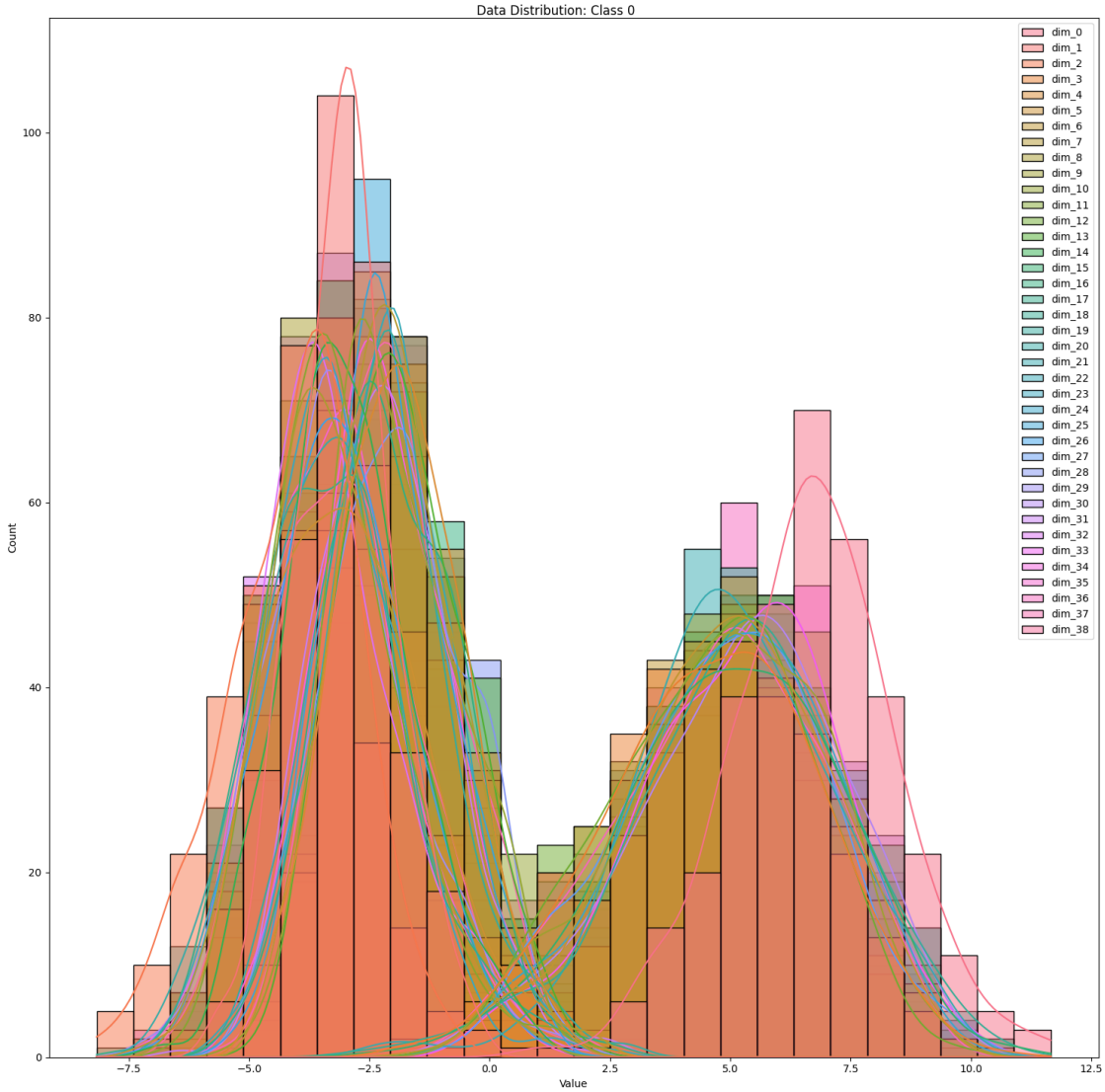


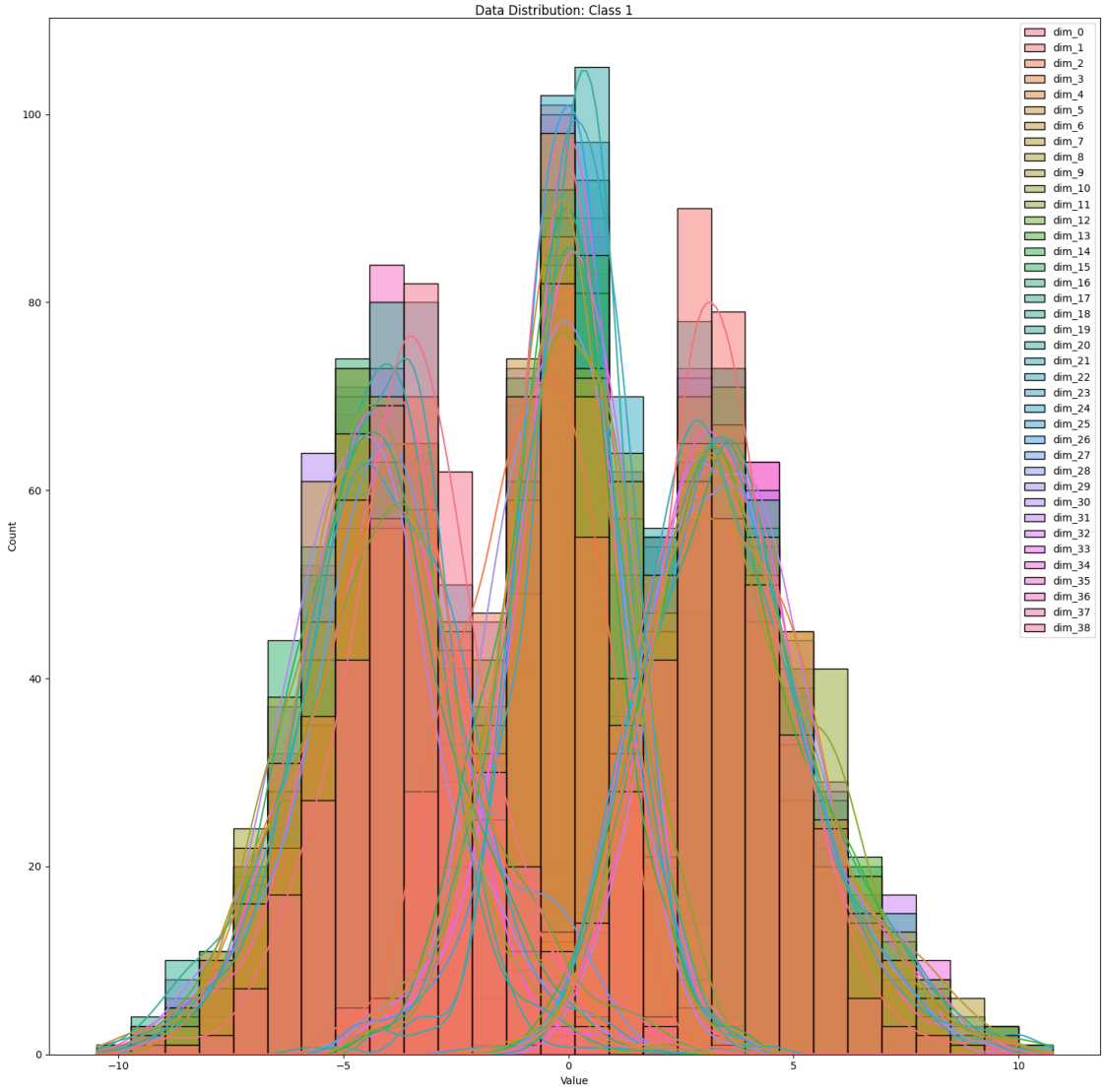

```
Test visualize
```

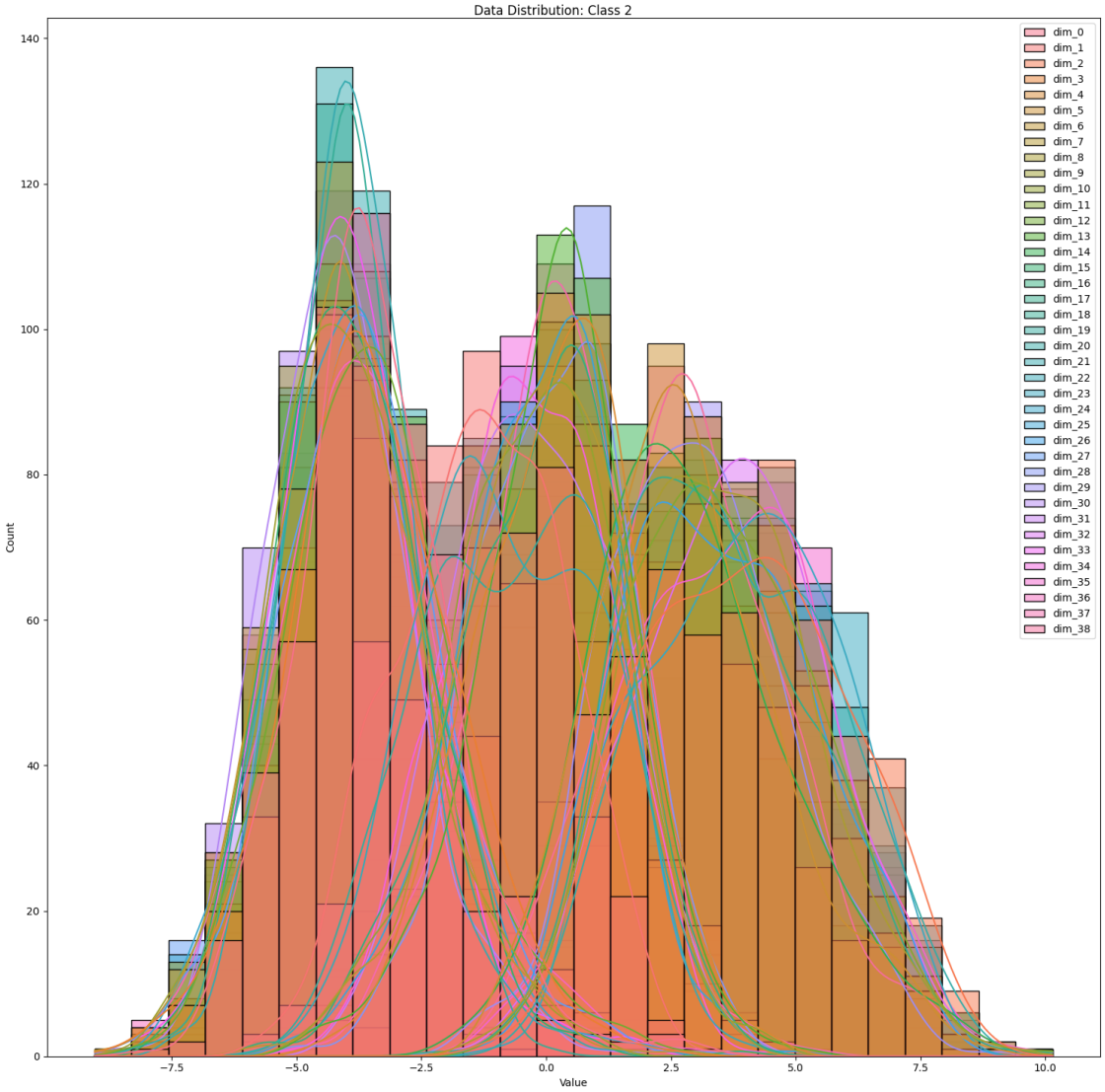


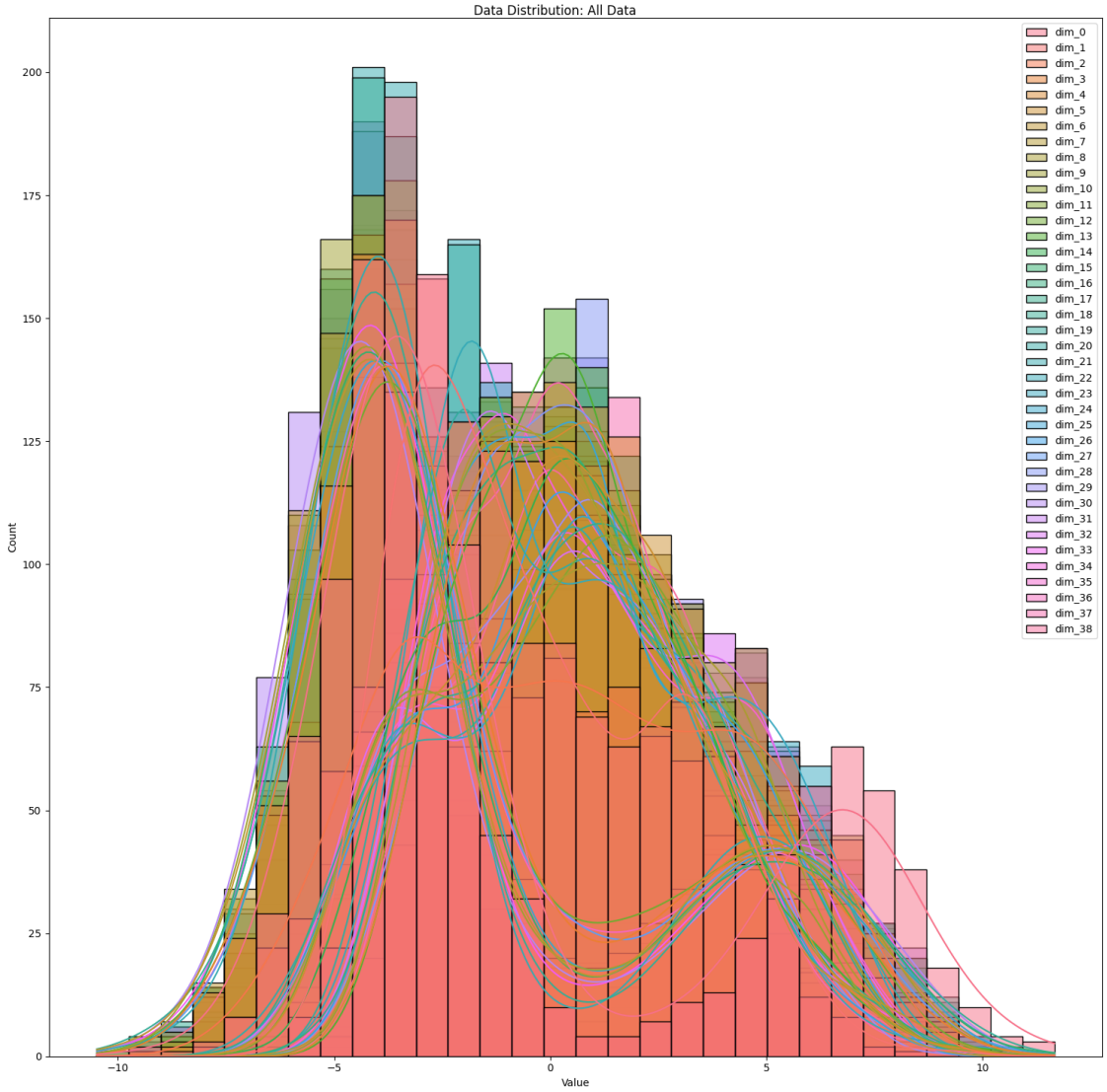
[ ]:
```
print("Train visualize")
visualize_distribution_histogram(train_features_39, train_labels_39)
print("Validation visualize")
visualize_distribution_histogram(valid_features_39, valid_labels_39)
print("Test visualize")
visualize_distribution_histogram(test_features_39, test_labels_39)
```

```
Train visualize
```

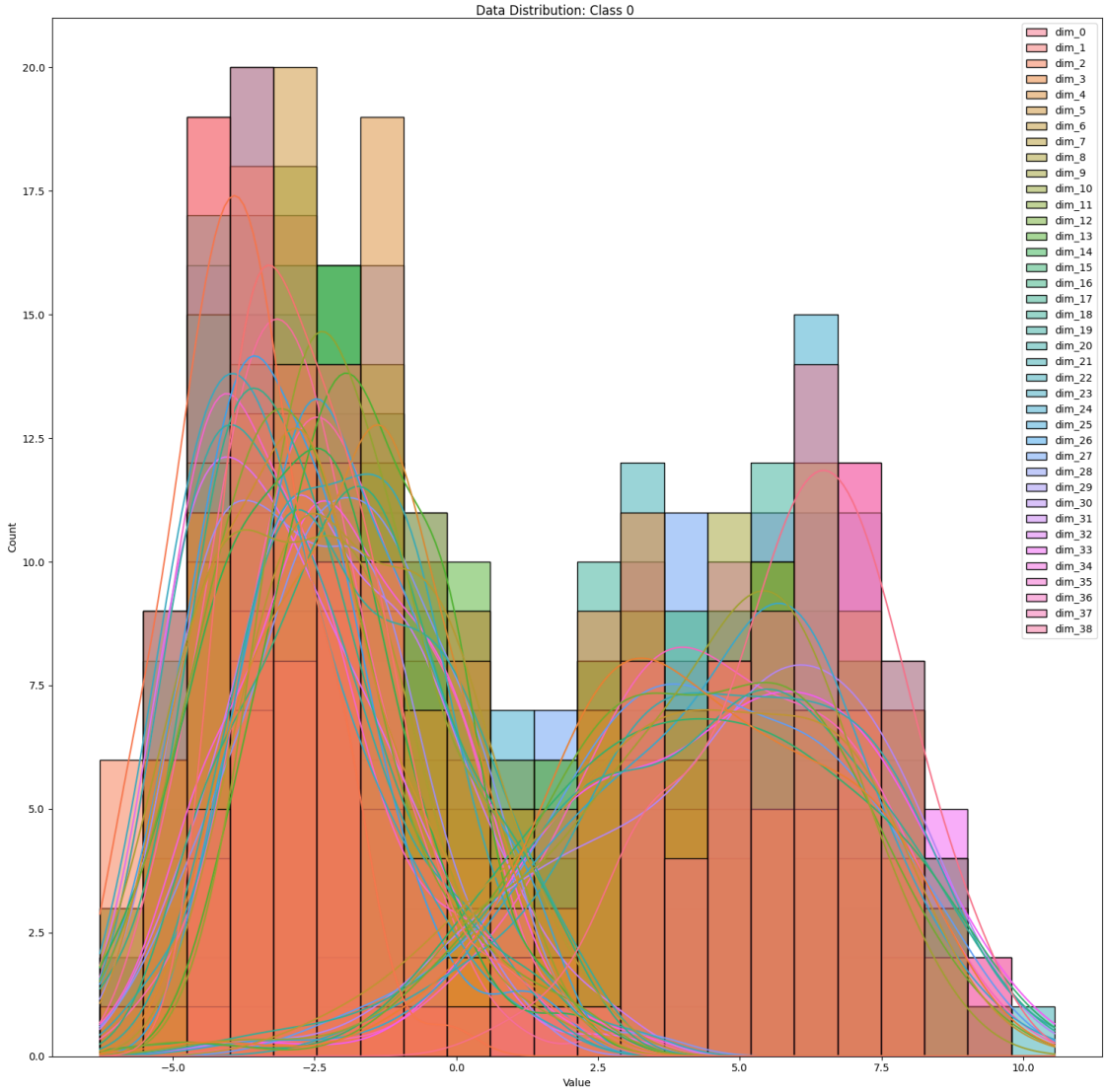

Data Distribution: Class 0
dim_0
dim_1
dim_2
dim_3
dim_4
dim_5
dim_6
dim_7
dim_8
dim_9
dim_10
dim_11
dim_12
dim_13
dim_14
dim_15
dim_16
dim_17
dim_18
dim_19
dim_20
dim_21
dim_22
dim_23
dim_24
dim_25
dim_26
dim_27
dim_28
dim_29
dim_30
dim_31
dim_32
dim_33
dim_34
dim_35
dim_36
dim_37
dim_38
Count
0
20
40
60
80
100
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0
12.5
Value

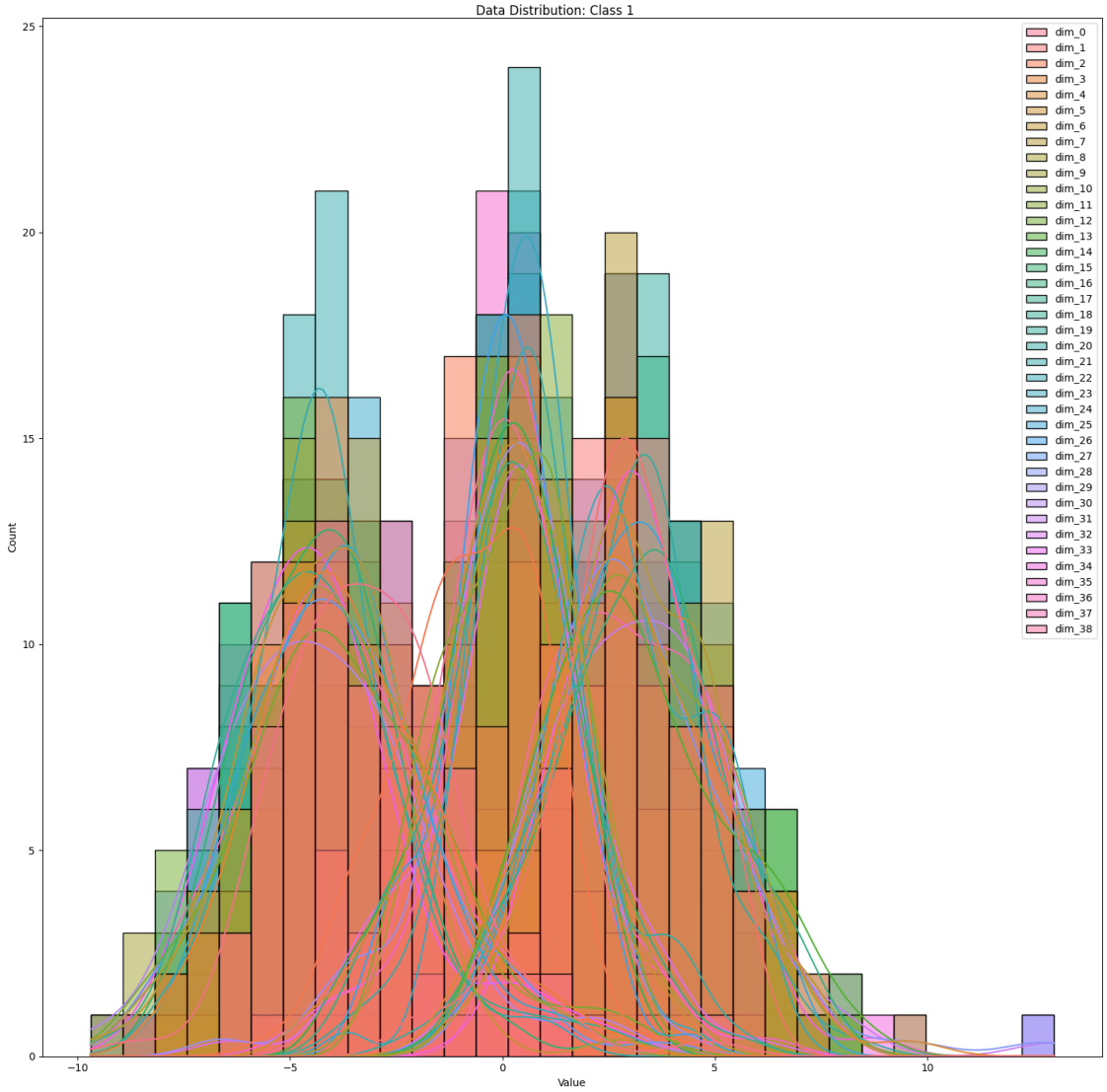

Data Distribution: Class 1
Count
Value
0
20
40
60
80
100
−10
−5
0
5
10
dim_0
dim_1
dim_2
dim_3
dim_4
dim_5
dim_6
dim_7
dim_8
dim_9
dim_10
dim_11
dim_12
dim_13
dim_14
dim_15
dim_16
dim_17
dim_18
dim_19
dim_20
dim_21
dim_22
dim_23
dim_24
dim_25
dim_26
dim_27
dim_28
dim_29
dim_30
dim_31
dim_32
dim_33
dim_34
dim_35
dim_36
dim_37
dim_38

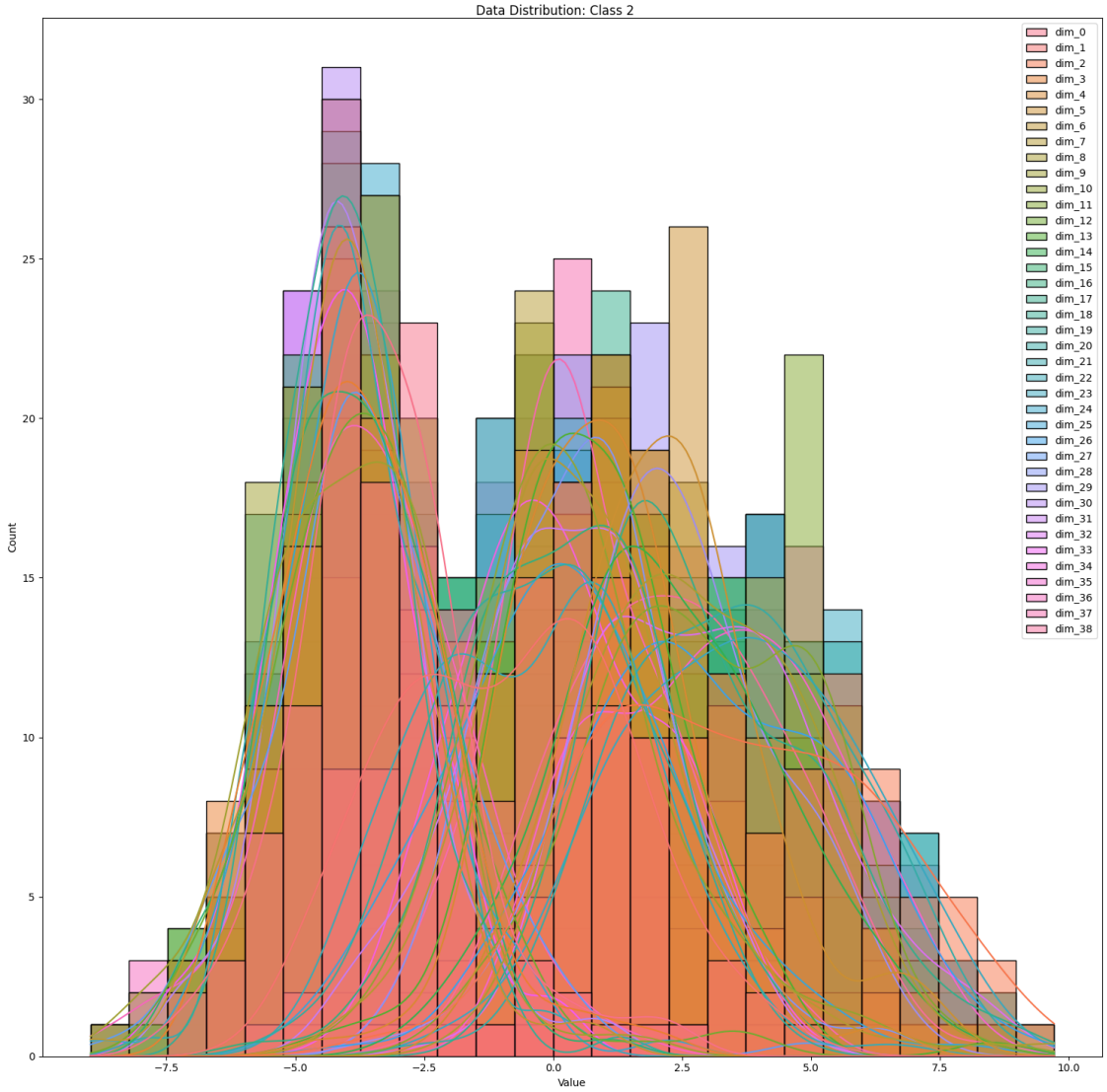

Data Distribution: Class 2
Count
Value
0
20
40
60
80
100
120
140
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0
dim_0
dim_1
dim_2
dim_3
dim_4
dim_5
dim_6
dim_7
dim_8
dim_9
dim_10
dim_11
dim_12
dim_13
dim_14
dim_15
dim_16
dim_17
dim_18
dim_19
dim_20
dim_21
dim_22
dim_23
dim_24
dim_25
dim_26
dim_27
dim_28
dim_29
dim_30
dim_31
dim_32
dim_33
dim_34
dim_35
dim_36
dim_37
dim_38

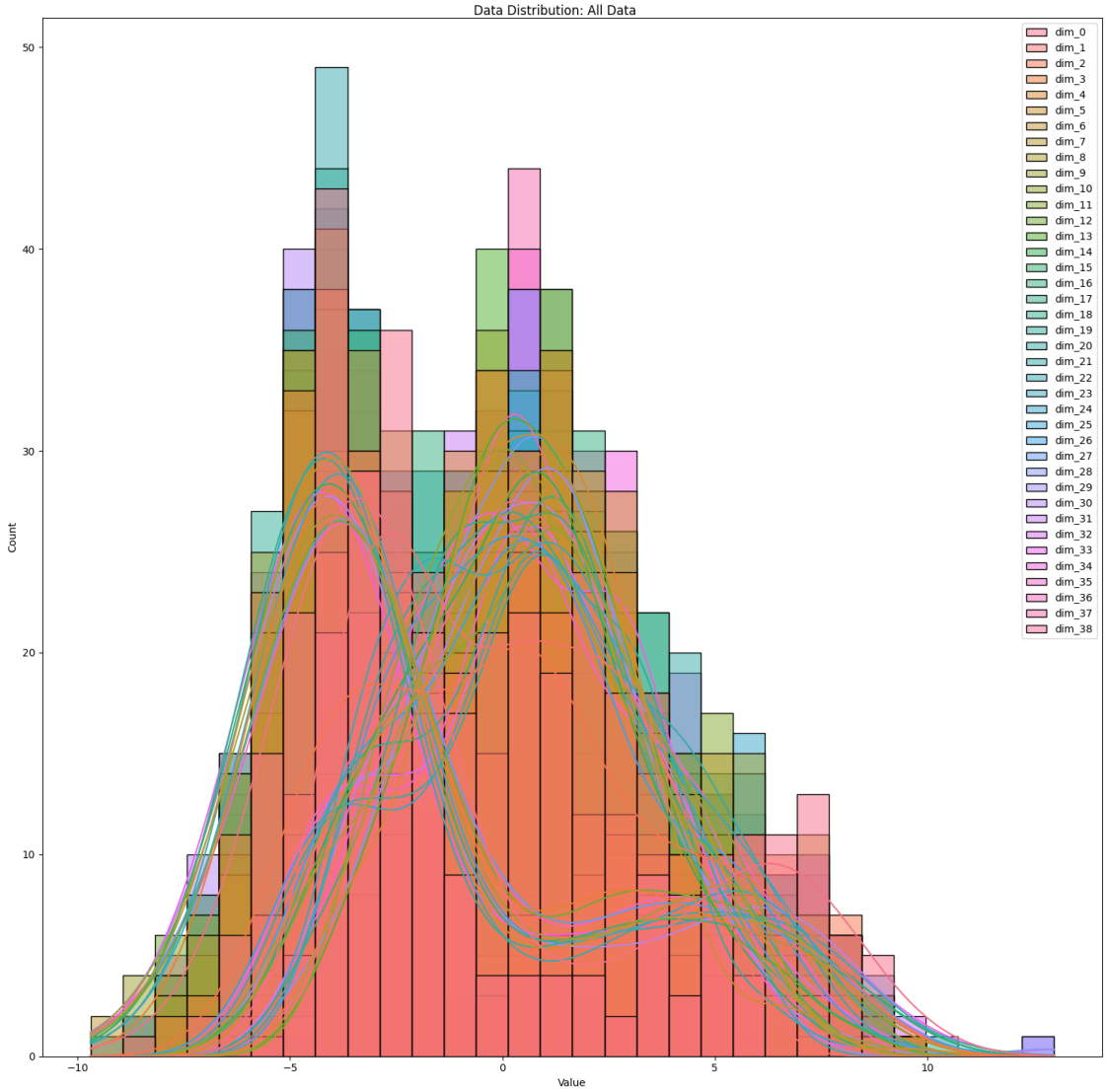


Validation visualize

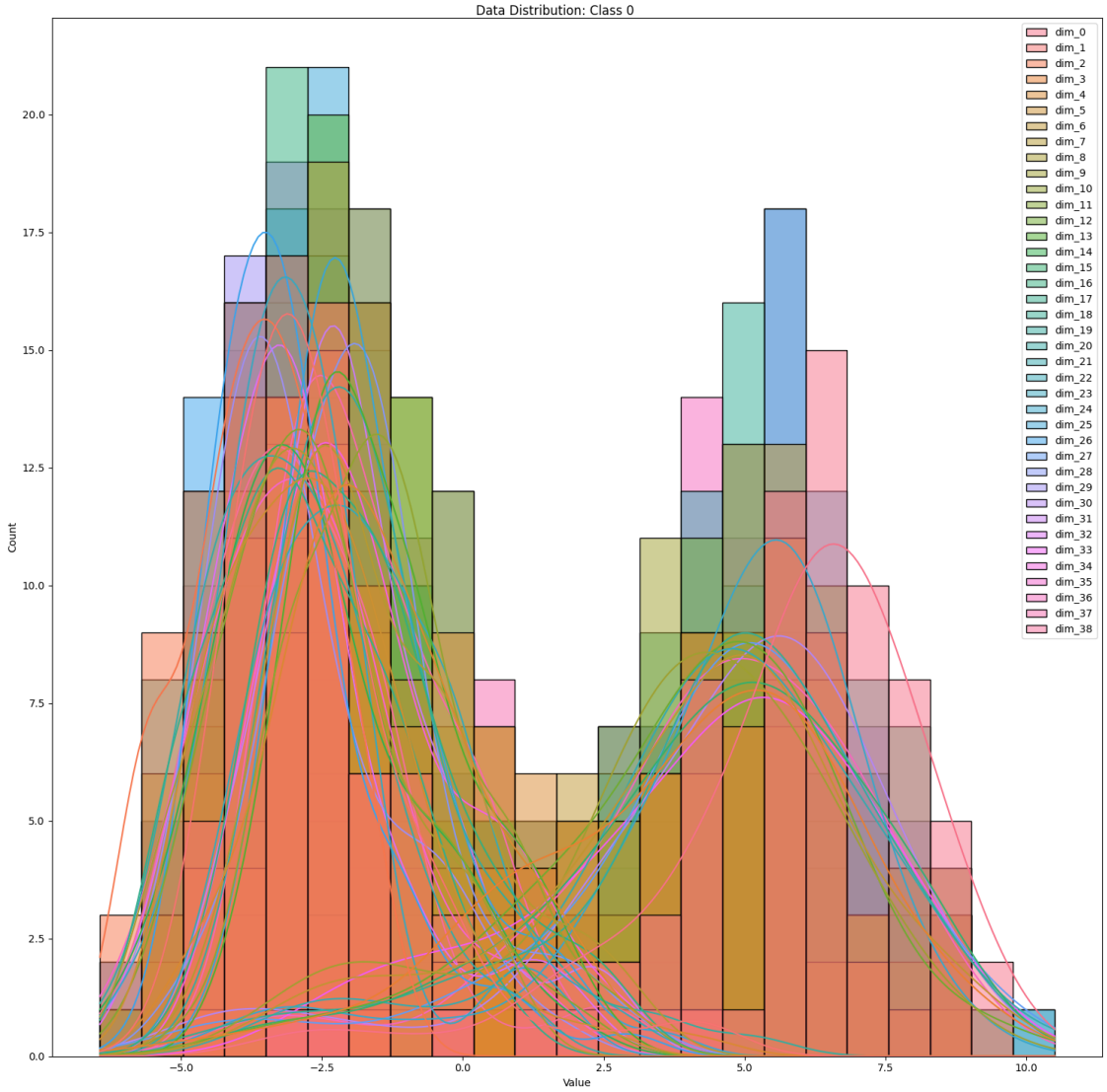

Data Distribution: Class 0
Count
Value
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0
dim_0
dim_1
dim_2
dim_3
dim_4
dim_5
dim_6
dim_7
dim_8
dim_9
dim_10
dim_11
dim_12
dim_13
dim_14
dim_15
dim_16
dim_17
dim_18
dim_19
dim_20
dim_21
dim_22
dim_23
dim_24
dim_25
dim_26
dim_27
dim_28
dim_29
dim_30
dim_31
dim_32
dim_33
dim_34
dim_35
dim_36
dim_37
dim_38

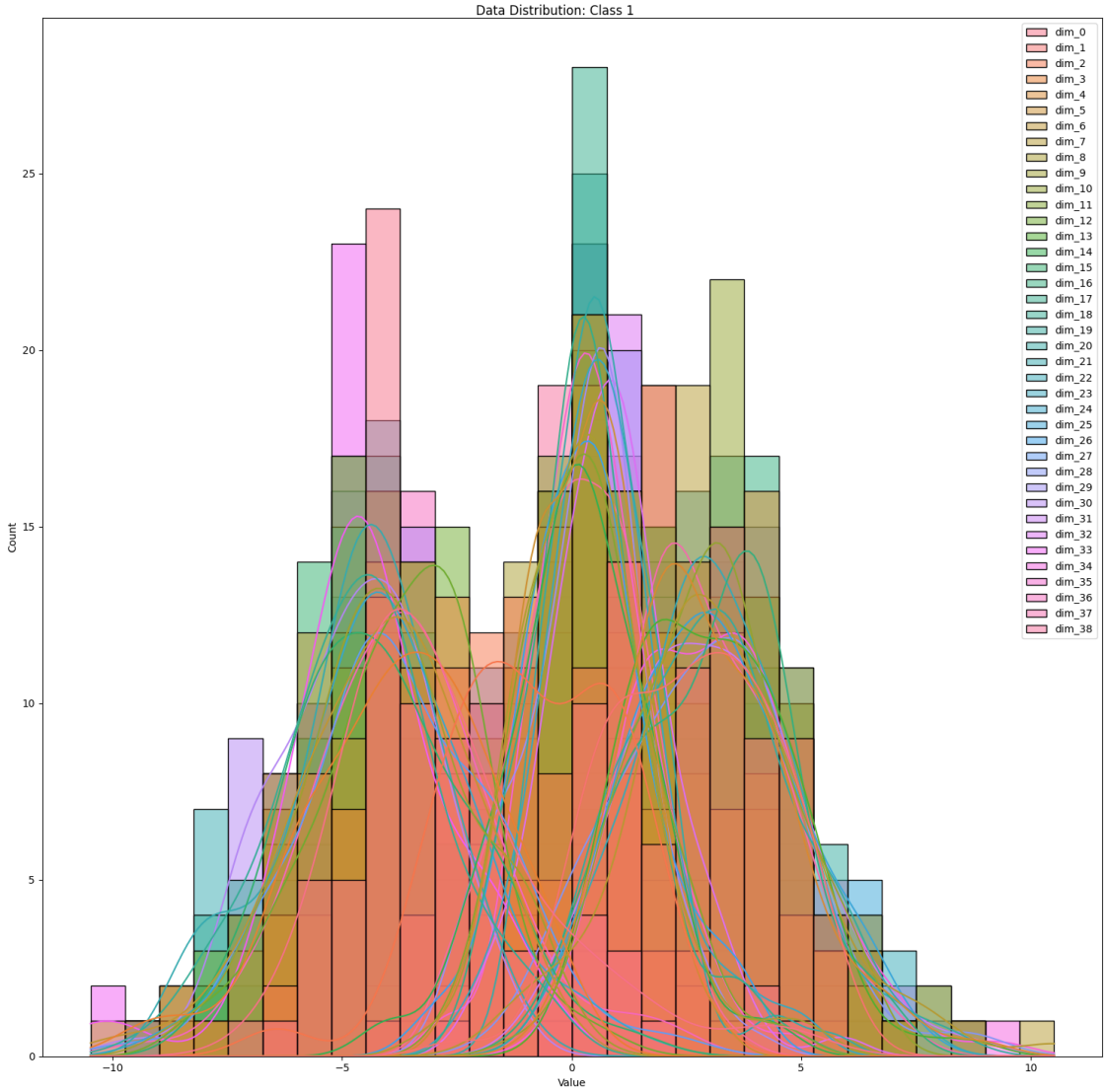

Data Distribution: Class 1
Count
Value
0
5
10
15
20
25
−10
−5
0
5
10
dim_0
dim_1
dim_2
dim_3
dim_4
dim_5
dim_6
dim_7
dim_8
dim_9
dim_10
dim_11
dim_12
dim_13
dim_14
dim_15
dim_16
dim_17
dim_18
dim_19
dim_20
dim_21
dim_22
dim_23
dim_24
dim_25
dim_26
dim_27
dim_28
dim_29
dim_30
dim_31
dim_32
dim_33
dim_34
dim_35
dim_36
dim_37
dim_38

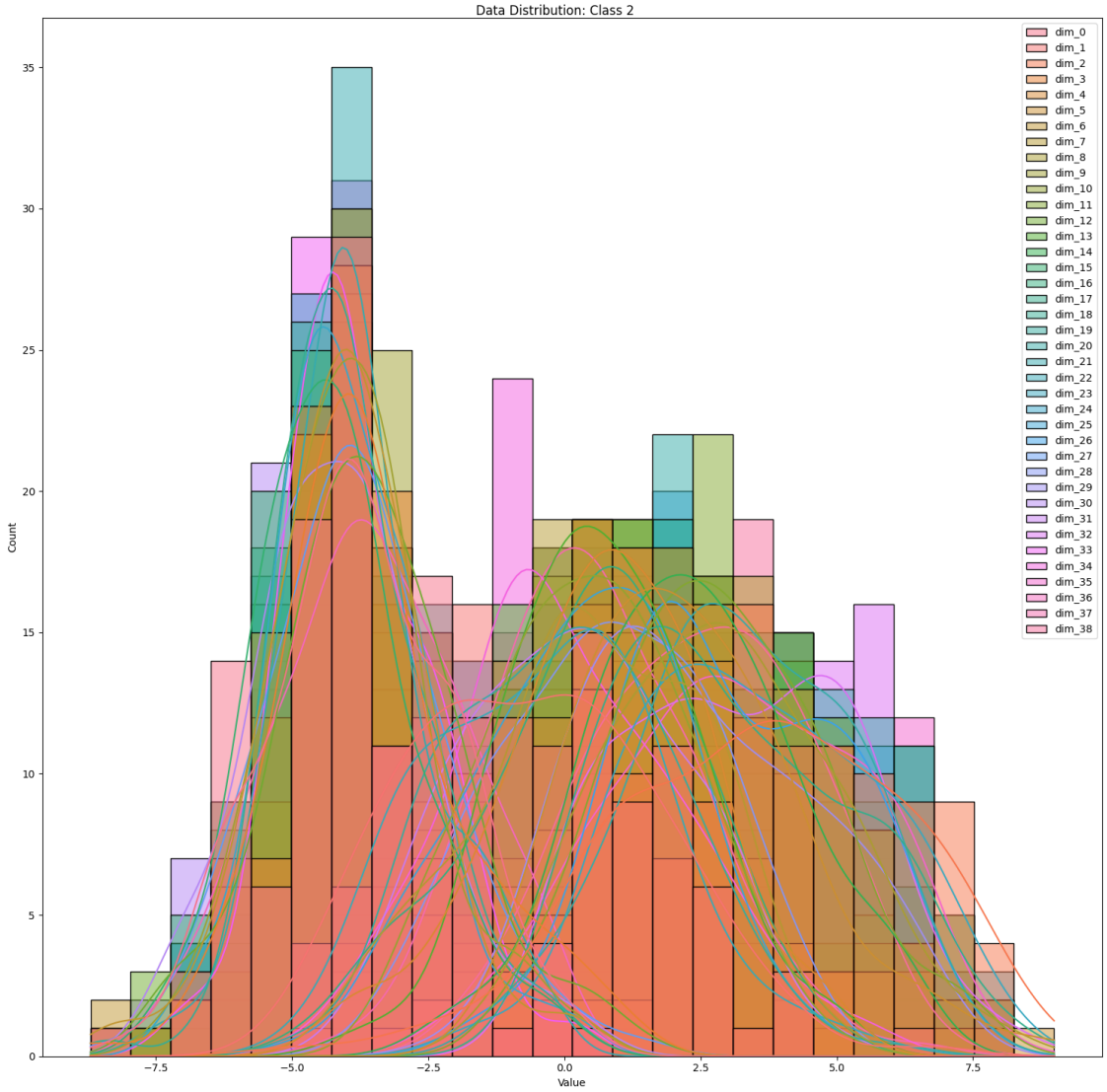

Data Distribution: Class 2
Count
Value
0
5
10
15
20
25
30
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0
dim_0
dim_1
dim_2
dim_3
dim_4
dim_5
dim_6
dim_7
dim_8
dim_9
dim_10
dim_11
dim_12
dim_13
dim_14
dim_15
dim_16
dim_17
dim_18
dim_19
dim_20
dim_21
dim_22
dim_23
dim_24
dim_25
dim_26
dim_27
dim_28
dim_29
dim_30
dim_31
dim_32
dim_33
dim_34
dim_35
dim_36
dim_37
dim_38

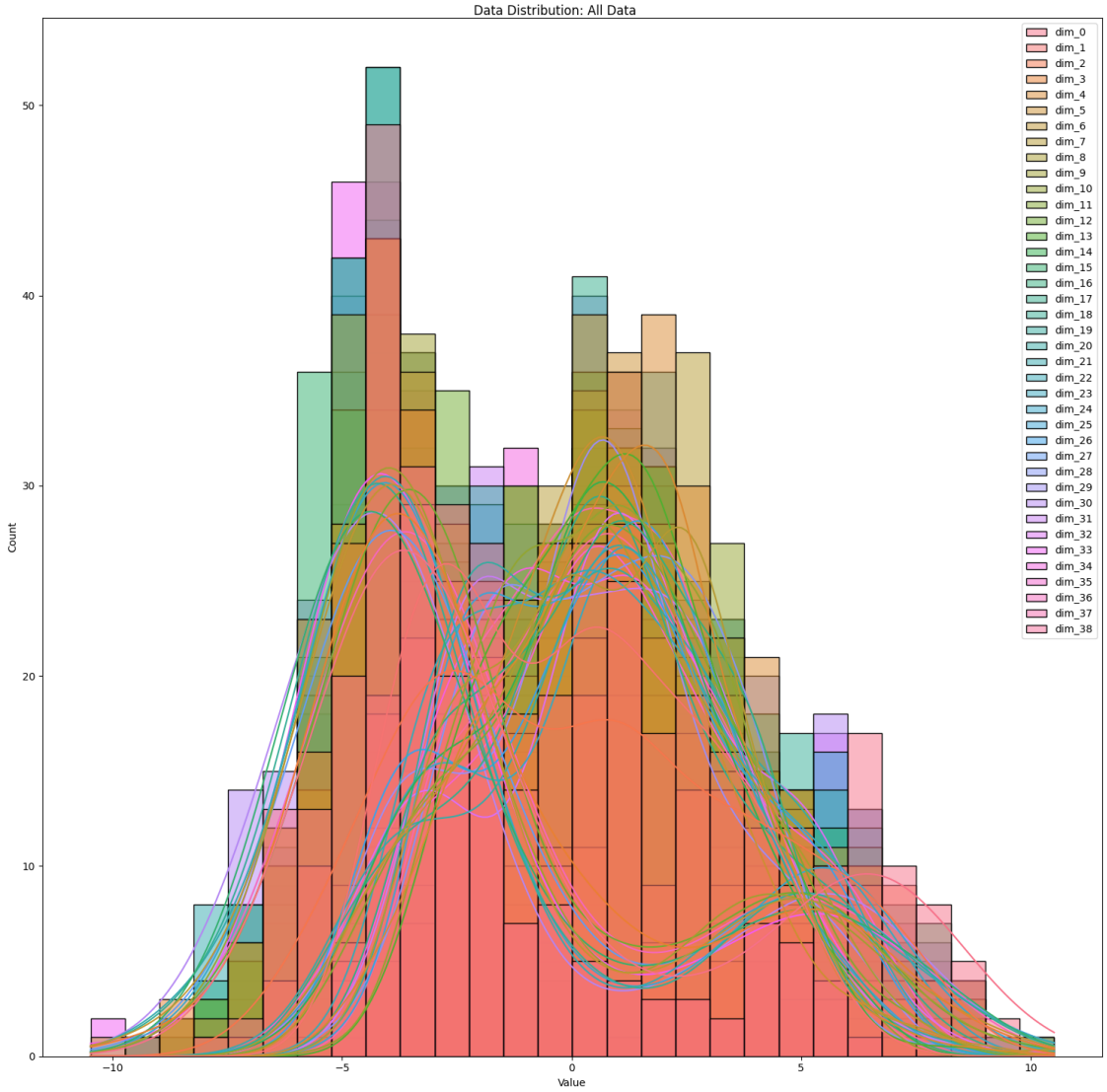


Test visualize


Data Distribution: Class 0
Count
Value
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0
dim_0
dim_1
dim_2
dim_3
dim_4
dim_5
dim_6
dim_7
dim_8
dim_9
dim_10
dim_11
dim_12
dim_13
dim_14
dim_15
dim_16
dim_17
dim_18
dim_19
dim_20
dim_21
dim_22
dim_23
dim_24
dim_25
dim_26
dim_27
dim_28
dim_29
dim_30
dim_31
dim_32
dim_33
dim_34
dim_35
dim_36
dim_37
dim_38


Data Distribution: Class 1
Count
Value
0
5
10
15
20
25
−10
−5
0
5
10
dim_0
dim_1
dim_2
dim_3
dim_4
dim_5
dim_6
dim_7
dim_8
dim_9
dim_10
dim_11
dim_12
dim_13
dim_14
dim_15
dim_16
dim_17
dim_18
dim_19
dim_20
dim_21
dim_22
dim_23
dim_24
dim_25
dim_26
dim_27
dim_28
dim_29
dim_30
dim_31
dim_32
dim_33
dim_34
dim_35
dim_36
dim_37
dim_38


Data Distribution: Class 2
Count
Value
0
5
10
15
20
25
30
35
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
dim_0
dim_1
dim_2
dim_3
dim_4
dim_5
dim_6
dim_7
dim_8
dim_9
dim_10
dim_11
dim_12
dim_13
dim_14
dim_15
dim_16
dim_17
dim_18
dim_19
dim_20
dim_21
dim_22
dim_23
dim_24
dim_25
dim_26
dim_27
dim_28
dim_29
dim_30
dim_31
dim_32
dim_33
dim_34
dim_35
dim_36
dim_37
dim_38


Data Distribution: All Data
Count
Value
0
10
20
30
40
50
−10
−5
0
5
10
dim_0
dim_1
dim_2
dim_3
dim_4
dim_5
dim_6
dim_7
dim_8
dim_9
dim_10
dim_11
dim_12
dim_13
dim_14
dim_15
dim_16
dim_17
dim_18
dim_19
dim_20
dim_21
dim_22
dim_23
dim_24
dim_25
dim_26
dim_27
dim_28
dim_29
dim_30
dim_31
dim_32
dim_33
dim_34
dim_35
dim_36
dim_37
dim_38