Name: Krushikesh Shelar Class: D15C Roll No: 51

Experiment 3

Aim: Perform Data Modeling on the dataset.

Theory:

Partition the dataset, ensuring that 75% of the records are included in the training dataset and 25% in the test dataset.

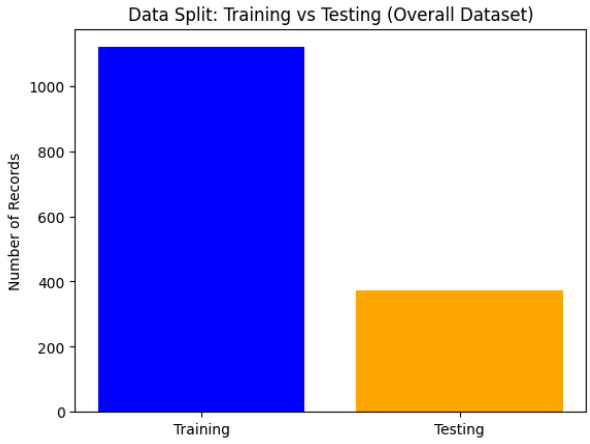
In this experiment, we partitioned a dataset of fleet data into a training set (75%) and a test set (25%) and validated this partitioning using a two-sample Z-test. The dataset consists of 1,492 records with features such as vehicle ID, fleet type, engine performance, and maintenance history.

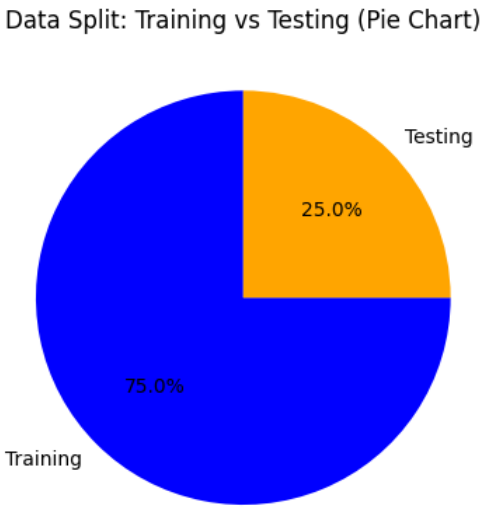
```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
df = pd.read_csv(list(uploaded.keys())[0])
train_data, test_data = train_test_split(df, test_size=0.25, random_state=42)
labels = ['Training', 'Testing']
sizes = [len(train_data), len(test_data)]
plt.bar(labels, sizes, color=['blue', 'orange'])
plt.title("Data Split: Training vs Testing (Overall Dataset)")
plt.ylabel("Number of Records")
plt.show()
plt.pie(sizes, labels=labels, autopct='%1.1f%%', colors=['blue', 'orange'], startangle=90)
plt.title("Data Split: Training vs Testing (Pie Chart)")
plt.show()
print("Total records in the training data set:", len(train_data))
print("Total records in the testing data set:", len(test_data))
```

Visualizing Data Partitioning: To confirm the proportions of the data split into training and test sets, we used a bar graph and a pie chart:

- **Bar Graph:** Displays the number of records in the training and test sets, visually confirming the 75%-25% split. The x-axis represents the two sets, and the y-axis shows their respective record counts.
- **Pie Chart:** Visually illustrates the percentage split between the training (75%) and test (25%) sets, providing an easy confirmation of the partition.

Name: Krushikesh Shelar Class: D15C Roll No: 51





Identifying the Total Number of Records in the Training Data Set: We used the train_test_split method to split the dataset into training and test sets. The partition was visually confirmed through a bar plot, showing the expected 75%-25% split, with 1,119 records in the training set and 373 in the test set.

```
Total records in the training data set: 1119
Total records in the testing data set: 373
```

Validation Using a Two-Sample Z-Test:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

```
[ ] # Perform Z-test on 'Total' column between train and test sets
    z_stat, p_value = ztest(train_df["Total"], test_df["Total"])

    print(f"Z-Statistic: {z_stat:.4f}")
    print(f"P-Value: {p_value:.4f}")

    # Interpretation
    alpha = 0.05
    if p_value < alpha:
        print("Reject the null hypothesis: The distributions are significantly different.")
    else:
        print("Fail to reject the null hypothesis: The distributions are similar.")

Z-Statistic: 2.0972
P-Value: 0.0360
Reject the null hypothesis: The distributions are significantly different.
```

To validate the partitioning, we performed a two-sample Z-test manually to compare the "Total" values between the training and test datasets. The Z-statistic was calculated based on the means, standard deviations, and sample sizes of both datasets.

The calculated Z-statistic was 2.0972, and the corresponding p-value was 0.0360. Since the p-value was less than the chosen significance level of 0.05, we rejected the null hypothesis, concluding that the distributions of the "Total" values in the training and test sets are significantly different.

Conclusion: The partitioning of the dataset into training and test sets was validated successfully. The Z-test indicated a significant difference between the datasets, suggesting that the partition may not be entirely reliable for further analysis. Additional checks or adjustments to the partitioning process might be necessary.