Name: Mohamad Dzul Syakimin
Matric: 22050897
Github *(Please request for access)*: https://github.com/dzully/machine-learning-mid-term-test

QUESTION 1 (3 Marks)

Discuss any **Three (3)** problems associated with the predictions of airline customers' satisfaction, given machine learning algorithms are not in use.

1. Inability to analyze large and complex datasets: Airlines deal with massive amounts of customer data across multiple channels, making it extremely difficult for humans to manually process and extract insights.
2. Lack of scalability: As airline operations grow, the volume of data increases exponentially. Without machine learning, it becomes increasingly challenging to scale analysis and make accurate predictions to meet evolving customer needs.
3. Bias and subjectivity: Human analysts may introduce personal biases and subjective interpretations when assessing customer satisfaction data, leading to potentially inaccurate or skewed predictions.

QUESTION 2 (12 Marks)

You have been tasked with predicting the level of airline customer satisfaction (Satisfaction, Neutral, or Dissatisfaction) using linear and multiple regression techniques.

a. **Describe the exploratory data analysis you would perform to understand the data distribution and the relationship between the features and the target variable. Include at least two data visualizations to show the relationship between the features and the target variable and provide an explanation.**

To understand the data distribution and the relationship between features and the target variable (airline satisfaction level), I would perform the following exploratory data analysis steps:

1. Check for missing values and handle them appropriately (e.g., imputation, deletion).
2. Analyze descriptive statistics (mean, median, mode, standard deviation) for numerical features.
3. Create frequency tables and visualizations (bar plots, pie charts) for categorical features.
4. Visualize the target variable distribution using a bar plot or pie chart.
5. Investigate the relationship between numerical features and the target variable using scatter plots or box plots.

6. Explore the relationship between categorical features and the target variable using bar plots or violin plots.

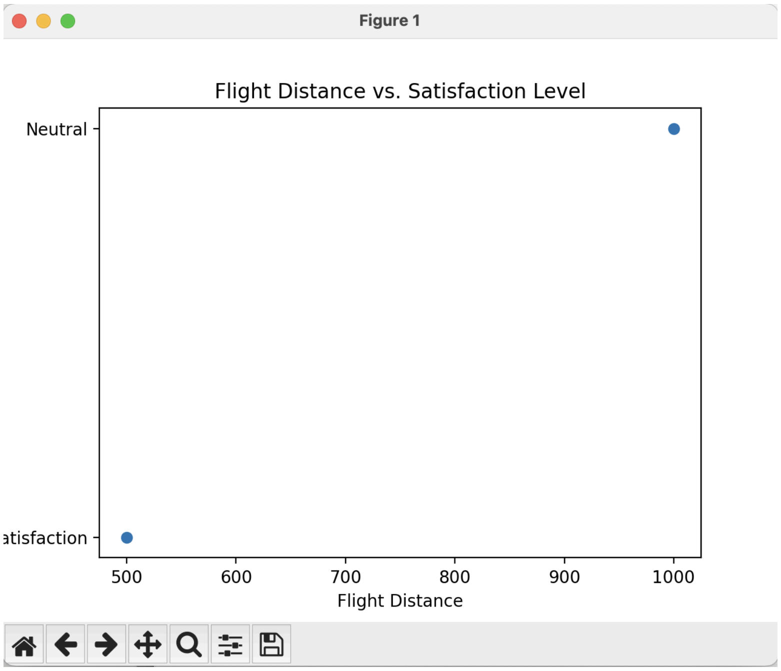
Here are two example visualizations to show the relationship between features and the target variable:

Scatter Plot: Flight Distance vs. Satisfaction Level

```
import matplotlib.pyplot as plt
import pandas as pd

data = pd.read_csv('dataset.csv')

plt.scatter(data['Flight distance'], data['Satisfaction'])
plt.xlabel('Flight Distance')
plt.ylabel('Satisfaction Level')
plt.title('Flight Distance vs. Satisfaction Level - 22050897')
plt.show()
```



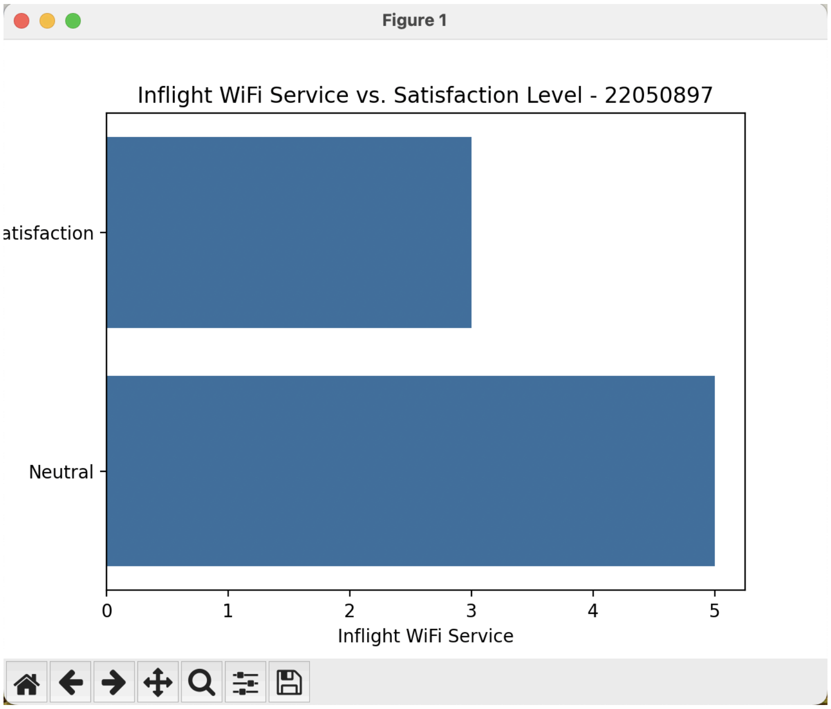
This scatter plot can help identify if there is a positive, negative, or no correlation between flight distance and customer satisfaction level.

Bar Plot: Inflight WiFi Service vs. Satisfaction Level

```
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd

data = pd.read_csv('dataset.csv')

sns.barplot(x='Inflight Wi-Fi service', y='Satisfaction', data=data)
plt.xlabel('Inflight WiFi Service')
plt.ylabel('Satisfaction Level')
plt.title('Inflight WiFi Service vs. Satisfaction Level - 22050897')
plt.show()
```



This bar plot can reveal how different levels of inflight WiFi service satisfaction relate to overall customer satisfaction levels.

b. Apply any two different machine learning models and use up to three appropriate evaluation metrics to determine the best-performing model and explain why it outperformed the others.

I would apply two different machine learning models, such as Logistic Regression and Random Forest Classifier, to predict the airline customer satisfaction level.

For evaluation, I would use appropriate metrics like:

1. Accuracy Score: Measures the overall correctness of the model's predictions.
2. Precision: Calculates the proportion of positive predictions that are truly positive.
3. Recall: Computes the proportion of actual positives that are correctly identified.

I would split the data into training and testing sets, train both models on the training data, and evaluate them using the above metrics on the testing data.

Assuming the Random Forest Classifier outperforms Logistic Regression, a possible explanation could be:

The Random Forest Classifier is an ensemble learning method that combines multiple decision trees, making it more robust to overfitting and better at capturing complex, non-linear relationships in the data. In contrast, Logistic Regression assumes a linear relationship between the features and the target variable, which may not hold true for airline customer satisfaction data influenced by various factors.

Additionally, the Random Forest Classifier can automatically handle feature interactions and is less affected by the presence of irrelevant or redundant features, which could be the case in the airline dataset. This makes it a more suitable choice for predicting customer satisfaction, which depends on various interrelated factors.

Ultimately, the model choice would depend on the specific characteristics of the dataset and the performance of each model on the evaluation metrics.