**Visualization and Analysis of Spam Emails**

Performed by **:** Bacem Boukhatem & Amir Amemi

1. Introduction:

* Technical Sheet Name: Spam Email Visualization and Analysis
* Description: This technical sheet demonstrates the visualization and analysis of spam emails using various tools and techniques.

2. Tools Used for Visualization and Analysis:

* Python Libraries: pandas, scikit-learn, nltk, Matplotlib, Seaborn, Plotly, WordCloud, keras
* Machine Learning Model: Logistic Regression

3. Data Loading and Exploration:

* Loaded the dataset from "Spam\_Data.csv" using pandas.
* Explored the dataset's structure, information, and summary statistics.

4. Data Preprocessing:

* Converted the "Category" column values to numerical representation (0 for "spam" and 1 for "ham").

5. Data Visualization:

* Utilized Plotly Express to create a pie chart depicting the distribution of spam and ham emails.
* Visualized the top 10 most common words in spam and ham emails using bar charts.

6. Feature Engineering:

* Created new features: "nb\_char" (number of characters), "nb\_words" (number of words), "nb\_sentences" (number of sentences).

7. Correlation Analysis:

* Examined the correlation between "Category," "nb\_char," "nb\_words," and "nb\_sentences" using a heatmap.

8. Pair Plot and Box Plot:

* Generated a pair plot to visualize relationships between features.
* Created a box plot to compare the number of sentences in spam and ham emails.

9. Word Cloud:

* Constructed a word cloud to visualize the most common words in the entire dataset.

10. Distribution of Word Length:

* Plotted histograms to show the distribution of average word length in spam and ham emails.

11. 3D Scatter Plot:

* Employed Plotly to create a 3D scatter plot using the features "nb\_char," "nb\_words," and "nb\_sentences."

12. Modeling and Evaluation:

* Split the dataset into training and testing sets.
* Utilized TF-IDF vectorization for text data.
* Trained a Logistic Regression model and evaluated its accuracy on test data.
* Visualized the confusion matrix and ROC curve.

13. Model Evaluation Metrics:

* Calculated precision, recall, and F1 score for the logistic regression model.

14. Prediction Examples:

* Demonstrated predictions for sample spam and ham email messages.