**SMS SPAM DETECTION USING   
DATA SCIENCE & MACHINE LEARNING  
PROJECT**

Steps:

1. Data Cleaning

2. EDA

3. Text Pre processing

4. Model Building

5. Evaluation

6. Improvements

7. Website

8. Website Deployment

**Data Cleaning**

1. Dropped unnecessary columns
2. Renamed column names
3. Labelled for indexing
4. Checked any null value is present
5. Checked any duplicate values are present
6. Dropped duplicates

**EDA**

1. Make a Pie Chart of ham and spam
2. It shows the imbalance of data
3. Download the package **import nltk**{NLTK – Natural Language Tool Kit}
4. Using **nltk** , calculate the **“num\_characters”, “num\_words”, “num\_sentences”**
5. Describe the data completely.
6. Describe the data of each ham and spam.
7. Visualize data in various aspects   
   [in terms of number of words, sentences etc.]
8. Plot the data using the **sea born** plot
9. Find the correlation coefficient

**Correlation Coefficient:**

**Correlation** is a statistical measure that describes the **relationship between two variables**—how they move together.

**Key Points:**

* If **one variable increases and the other also increases**, they have a **positive correlation** (e.g., height and weight).
* If **one variable increases while the other decreases**, they have a **negative correlation** (e.g., more exercise, less body fat).
* If there is **no consistent pattern**, they have **no correlation** (e.g., shoe size and intelligence).

The **correlation coefficient** is a number that shows how **strongly** two things are related.

* It **ranges from -1 to 1**:
  + **+1** → Perfect **positive** relationship (when one increases, the other also increases).
  + **0** → No relationship.
  + **-1** → Perfect **negative** relationship (when one increases, the other decreases).