Machine Learning Workflow

1- Project Setup

* Understand Business Goals
* set a measurable goal
* Data Collection

2- Data Preparation

* Data Cleaning
* Exploratory Data Analysis
* Data Pre-processing
* Data Splitting

3- Modeling

* Model selection
* Model Training
* Model Evaluation (make predictions, calculate accuracy)

1. Project setup:

* Understand Business Goals: understand deeply the business goal behind the model being proposed. A deep understanding of your business goals will help you scope the necessary technical solution, data sources to be collected, how to evaluate model performance, and more.
* Set the Goal: Define the problem you want to solve with machine learning, Set clear goals and objectives for your model.
* *Data Collection:* Collect all the data you need for your models, whether from your own organization, public or paid sources.

1. Data Preparation:

First of all, you need to import your dataset (*pandas*), second you must understand the data by visualization ex: Show the first few rows to get the first ideas on your dataset.

* Data Cleaning:
* **Identification of the target and features**
* **Number of rows and columns:** allows you to assess whether the data is suitable for your machine learning task. You must have enough data to train an effective model
* **Types of variables:** Categorical variables may require encoding, while numeric variables may require normalization. Understanding variable types guides appropriate preprocessing.
* **Detection of missing or nulls values:** allows you to decide how to handle these missing data. You can choose between value imputation (for example, replacing missing values with the mean of existing values) or the removal of rows or columns containing missing data.
* **outlier analysis (using box plot for features):** Outliers can result from data entry errors, incorrect measurements, or other data quality issues. Identifying and managing these values helps improve data reliability.
* Exploratory Data Analysis:
* **Target visualization:** This visualization can help understand the distribution of the target variable, so we can use:
* **Histogram:** For a continuous target variable, you can create a histogram to visualize its distribution. This provides an overview of the frequency of different values and their spread.
* **Box Plot:** For a Discrete variable, A box plot represents the median, quartiles, and potential outliers of the target variable. It also displays the data distribution.
* **Visualizing the relationships between features and the target:**

