# 1 Introduction

As you have seen in **the introduction**, Machine Learning modeling **Workflow** consists of three main parts.

1. Data-Preprocessing
2. Machine Learning (ML) Modelling
3. Model Performance Accuracy Analysis

**Data preprocessing** is the first part of ML modeling work. Data must be pre-processed to make it clean and select the right features for the ML modeling.

**Data MUST be Clean**! The **Right Features** must be **Selected** for the Modeling.

Otherwise, **Garbage In** Results in **Garbage Out**. The model will not be useful.

The main DATA Pre-processing steps are:

1. Data Cleaning
2. Feature Selection
3. Data Smoothing/Filtering (Optional)

# 2 Practice

We will apply the man data preprocessing steps mentioned above. Here, we use **drilling data** for practice. (File name is **DrillingData.xlsx**)

**A) Data Cleaning**

1. Import important libraries to read and plot data in Python.
2. Explore About the Dataset

A1) Check if the dataset is float, Int, or Object (Data must be float for ML)

A2) Check Unrecorded data (blank/null) and remove it.

A3) Check Duplicate data and remove it.

A4) D**etect and remove outliners** based on Target data. Then, **save the File with the NEW File Name! Cleanoutlier.xlsx**

**B) Feature Selection**

B1) Import the **outlier removed** and cleaned data file, from **Step A4**

B2) Do **Feature Selection** based on **correlation analysis**. Then Save the Selected **Feature File with the New File Name**

**C) Data Filtering/ Smoothing**

C1) Import the Feature Selected File (**Step B2**).

C2) Apply a Smoothing filter if required and save the file with **the NEW file NAME**.

Finally, the saved, **Cleaned/Feature selected (Step B2)** or **Cleaned/Feature selected /Smoothed (Step C2)** FILE will be used for Machine Learning Modeling. Typically, it is better to use the **Step B2 saved file** since the filtered data may lose information from the dataset.

# **Project #3- Task #1: Data-Preprocessing (Data Cleaning and Feature Selection)**

Using your project file (**ProjectData2024.xlsx)**, apply the above data processing steps to make it clean and select the right features. You need to comment and follow the reporting presentation format. Remember to save the final cleaned, and the features selected file with **a NEW name**. It will be used for Machine learning modeling projects #2, #3, and #4.