**1. Are all the columns necessary or can we drop any?**

Not all columns may be necessary. After exploration, we can consider dropping:

* duration (: It's highly correlated with the output, but it's only known **after** the campaign.
* pdays: Contains many 999 values — could be dropped or transformed.
* default: Mostly 'no' or 'unknown' — may have low variance or value.
* Highly correlated numeric columns — we can drop one from a correlated pair (e.g., emp.var.rate and nr.employed).

We'll finalize which to drop during **data exploration and correlation analysis**.

**2. Does the data contain any issues?**

Yes, some issues typically found:

* 'unknown' values in several categorical columns like job, education, contact, poutcome, etc.
* Imbalanced target variable (y: many more 'no' than 'yes').
* Mixed types: numerical, ordinal, and nominal categorical variables.
* Some columns may have missed or uninformative values (pdays = 999).

These need to be cleaned/encoded appropriately.

**3. What ML task is this? Classification? Regression? Clustering?**

**Binary Classification**

We are predicting whether a client will **subscribe to a term deposit** (y: yes/no).

So it's a supervised learning task → **binary classification**.

**4. What are the data types of the columns? What pre-processing should you apply?**

**Categorical Columns:**

job, marital, education, contact, month, day\_of\_week, poutcome, etc.

Preprocessing: Use **Label Encoding** or **One-Hot Encoding**.

**Numerical Columns:**

* age, campaign, previous, emp.var.rate, euribor3m, etc.
* Preprocessing: Use **StandardScaler** or **MinMaxScaler** to normalize.

**Special Cases:**

* pdays: treat 999 as "not previously contacted" — consider binning or separate indicator column.
* y: needs to be converted to **binary (0/1)** for model training.

**5. What is the most appropriate metric for this model?**

Since the dataset is **imbalanced**, use metrics that account for class imbalance:

* **F1-Score** (most balanced between precision and recall)
* **ROC-AUC Score**
* **Accuracy** is not ideal alone — it can be misleading in imbalanced datasets