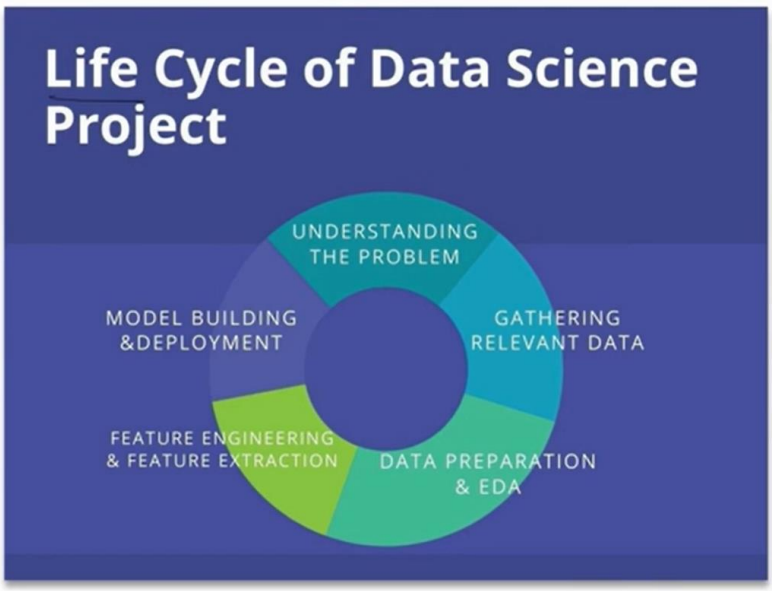
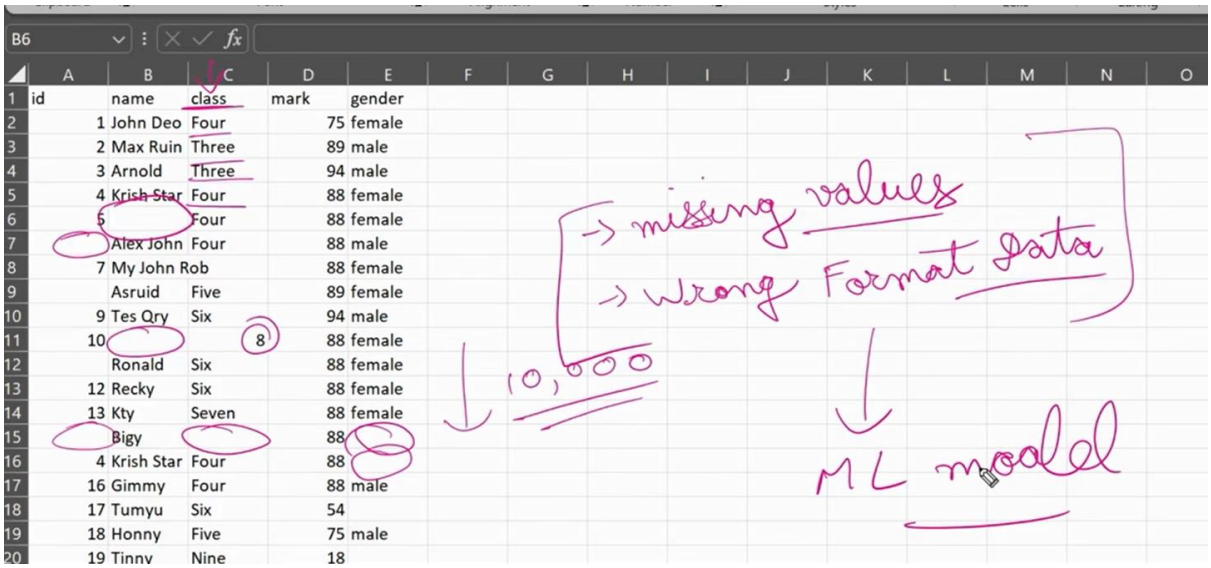
Introduction to the data science workflow

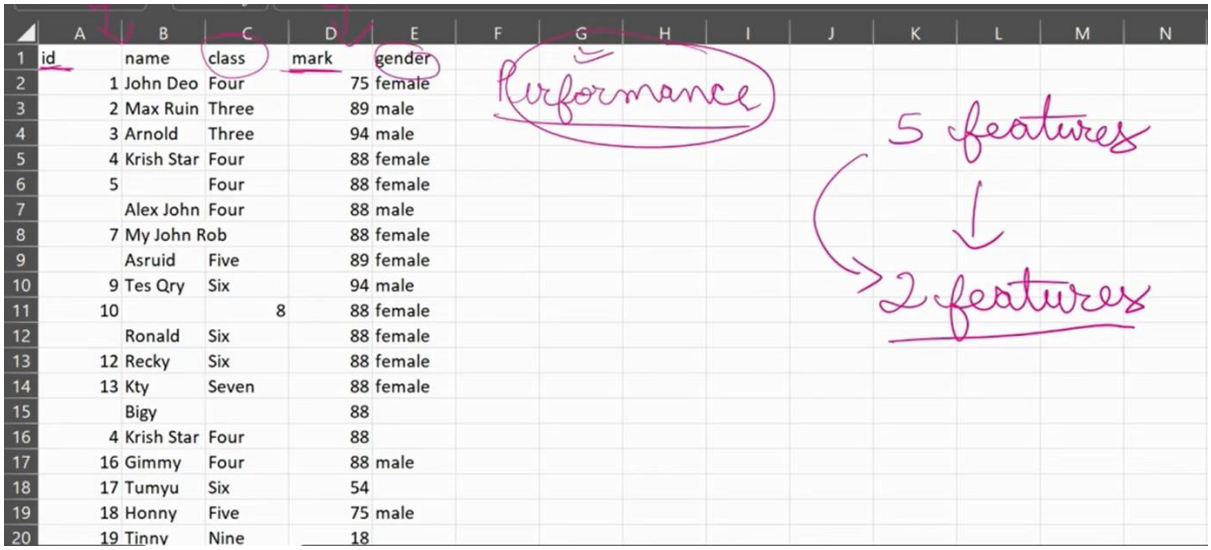


STEP 3: Preprocessing and Exploratory Data Analysis like handling the missing values and all :



id	name	class	mark	gender
1	John Deo	Four	75	female
2	Max Ruin	Three	89	male
3	Arnold	Three	94	male
4	Krish Star	Four	88	female
5		Four	88	female
	Alex John	Four	88	male
7	My John Rob		88	female
	Asruid	Five	89	female
9	Tes Qry	Six	94	male
10			8	88 female
	Ronald	Six	88	female
12	Recky	Six	88	female
13	Kty	Seven	88	female
	Bigy		88	
4	Krish Star	Four	88	
16	Gimmy	Four	88	male
17	Tumyu	Six	54	
18	Honny	Five	75	male
19	Tinny	Nine	18	

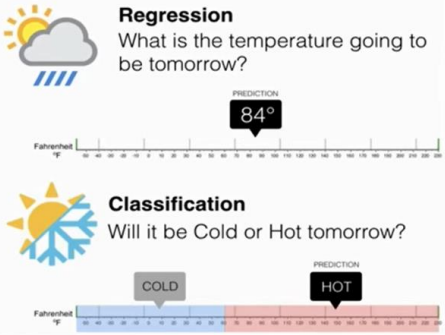
STEP 4 : Like finding the performance of the child we need only two features i.e. id and marks.
This is called feature selection.



id	name	class	mark	gender
1	John Deo	Four	75	female
2	Max Ruin	Three	89	male
3	Arnold	Three	94	male
4	Krish Star	Four	88	female
5		Four	88	female
	Alex John	Four	88	male
7	My John Rob		88	female
	Asruid	Five	89	female
9	Tes Qry	Six	94	male
10		8	88	female
	Ronald	Six	88	female
12	Recky	Six	88	female
13	Kty	Seven	88	female
	Bigy		88	
4	Krish Star	Four	88	
16	Gimmy	Four	88	male
17	Tumyu	Six	54	
18	Honny	Five	75	male
19	Tinny	Nine	18	

Understanding the Problem

- Define the objectives clearly.
- Determine how to measure the project's success.
- Identify the problem type (e.g., classification, regression).



Gathering Relevant Data

- Collect data from available sources.
- Ensure data is relevant to the problem.
- Consider the volume, variety, velocity, and veracity of the data.

Data Preparation & EDA (Exploratory Data Analysis)

- Clean the data (handle missing values, remove duplicates).
- Perform exploratory analysis to understand the data.
- Normalize or scale the data if necessary.

Feature Engineering & Feature Extraction

- Create new features that can help improve model performance.
- Reduce dimensionality if the feature space is too large.
- Select the most important features to be used for modeling.

Model Building & Deployment

- Choose appropriate algorithms and train models.
- Validate model performance using cross-validation.
- Deploy the model for real-time use or batch processing.