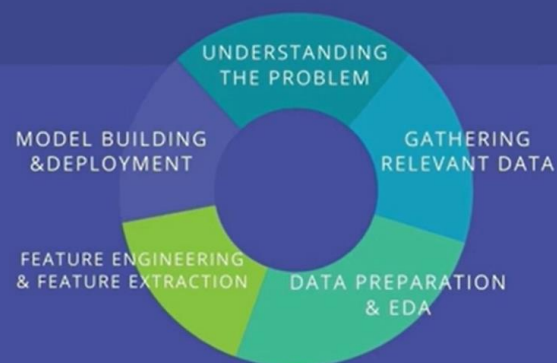


Introduction to the data science workflow

Life Cycle of Data Science Project



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

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	id	name	class	mark	gender										
2	1	John Deo	Four	75	female										
3	2	Max Ruin	Three	89	male										
4	3	Arnold	Three	94	male										
5	4	Krish Star	Four	88	female										
6	5	Alex John	Four	88	female										
7	7	My John Rob		88	female										
8		Asruid	Five	89	female										
9	9	Tes Qry	Six	94	male										
10	10	Ronald	Six	88	female										
11	12	Recky	Six	88	female										
12	13	Kty	Seven	88	female										
13	15	Bigy		88											
14	16	Gimmy	Four	88	male										
15	17	Tumyu	Six	54											
16	18	Honny	Five	75	male										
17	19	Tinnv	Nine	18											

Handwritten notes on the table:

- Annotations for data quality issues: "missing values" and "wrong Format data" with arrows pointing to the 'class' column (rows 7, 8, 13, 15).
- Annotation for a large value: "10,000" with an arrow pointing to the 'mark' column (row 11).
- Annotation for feature selection: "ML model" with an arrow pointing to the 'id' and 'mark' columns.

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

This is called feature selection.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	id	name	class	mark	gender									
2	1	John Deo	Four	75	female									
3	2	Max Ruin	Three	89	male									
4	3	Arnold	Three	94	male									
5	4	Krish Star	Four	88	female									
6	5	Alex John	Four	88	female									
7	7	My John Rob		88	female									
8		Asruid	Five	89	female									
9	9	Tes Qry	Six	94	male									
10	10	Ronald	Six	88	female									
11	12	Recky	Six	88	female									
12	13	Kty	Seven	88	female									
13	15	Bigy		88										
14	16	Gimmy	Four	88	male									
15	17	Tumyu	Six	54										
16	18	Honny	Five	75	male									
17	19	Tinnv	Nine	18										

Handwritten notes on the table:

- Annotation for feature selection: "Performance" with an arrow pointing to the 'mark' column.
- Annotation for feature selection: "5 features" with an arrow pointing to the 'id', 'name', 'class', 'mark', and 'gender' columns.
- Annotation for feature selection: "2 features" with an arrow pointing to the 'id' and 'mark' columns.

Understanding the Problem

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



Regression

What is the temperature going to be tomorrow?



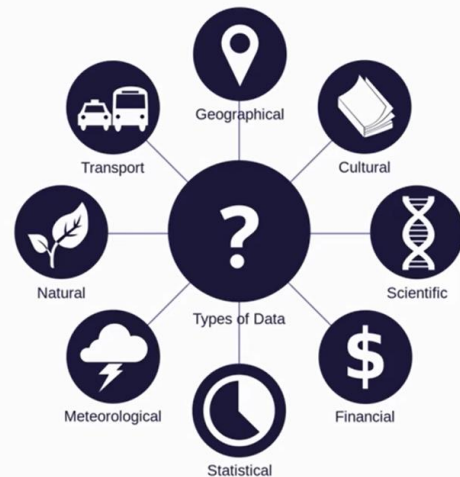
Classification

Will it be Cold or Hot tomorrow?



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.

