CRISP-DM (Cross Industry Standard Process for Data mining)

- standard process model for data science
- cyclical & iterative: after finishing one phase we revisit earlier phases to refine our work.

Phases of CRISP-DM

1. Business Understanding
2. Data understanding
3. Data preprocessing/preparation
4. Modeling
5. Evaluation
6. Deployment.

1) Business Understanding:

Understanding problem from business prespective and translating into data science.

Phases:-

(a) Determine business objective:

- understand what business tries to achieve
- eg. Reduce customer churn, Inc conversion r

(b) Access situation:

- identify stakeholders, resources, risk constraints

(c) Determine data science goals :

⇒ Translate business → data science
⇒ eg. build churn prediction, next month's sale predict.

(d) Produce Project Plan :

→ plan that outlines timelines, tasks, responsibil

2) Data Understanding :

Goal to collect initial data, explore it, access its quality

Phases :

(a) Collect Initial data :

→ Gather data from database, CSV files, API's, Web Scrapping, CRM tools.

(b) Describe data :

→ no. of rows & cols, data types, unique values, logically structuring of data

(c) Explore data :

→ visualise, relationship b/w cols, search pattern, EDA + visualization.

(d) Verify Data :

→ identify missing values, duplicates, invalid format (eg. age = -5), inconsistent values (like eg: 'USA', 'United State', 'us')

3. Data Preprocessing / Preparation

→ Construct final dataset from raw data. data is clean, relevant & right format

Phases:

(a) Data selection :

→ choose only relevant cols, drop unnecessary

(b) Data cleaning :

→ Handle missing values (mean, median, mode), outliers (zscore, IQR, cap-floor), duplicates, inconsistency

(c) Data Transform :

→ Encode category values (eg. one hot encoding)
→ scaling / normalization (zscore standard, MinMax)

(d) Feature Engineering :

→ build new features from existing data

(e) Data Integration :

→ combine multiple datasets into one (eg merge cust. detail and trans c. hist)

(f) Data Formatting :

→ ensure correct datatypes (eg. convert to datetime, string, int, float)

---

(4) Data Modeling :

→ choosing modeling techniques, train models

Phases :

(a) Select Modeling Techniques :

eg. predict if customer will churn.

(b) data splitting :

split into training, validation & test set.

(c) Build & Train :

→ use data to train models

(d) Tune model parameters :

→ grid, random, bayesian search

(5) Evaluate Model

→ final eval.

(5) Evaluation :

→ ensures last model fully solve business problem

Phase :

(a) Final Model Report
(b) Business Evaluation summary
(c) Model comparison table
(d) Risks & limitation report
(e) Go / No-Go decision

(6) Deploy :

→ Launch the app

Phases

(a) Deployment Plans
(b) Model or Insight delivery
(c) POCs
(d) Monitoring & maintenance
(e) User train & support
(f) Review & report