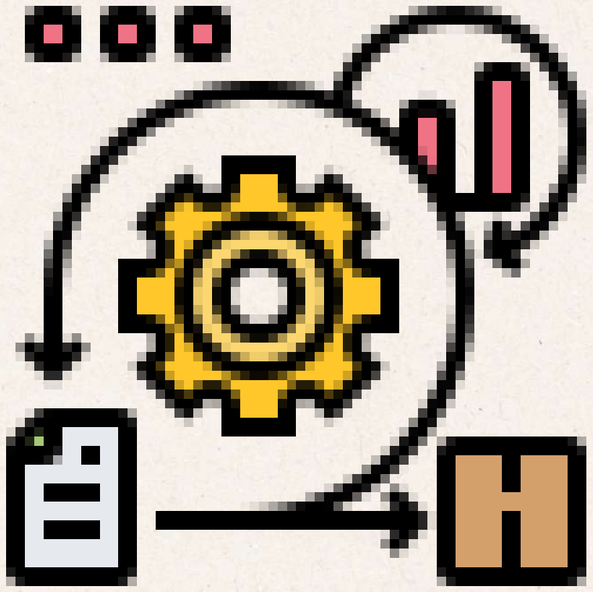
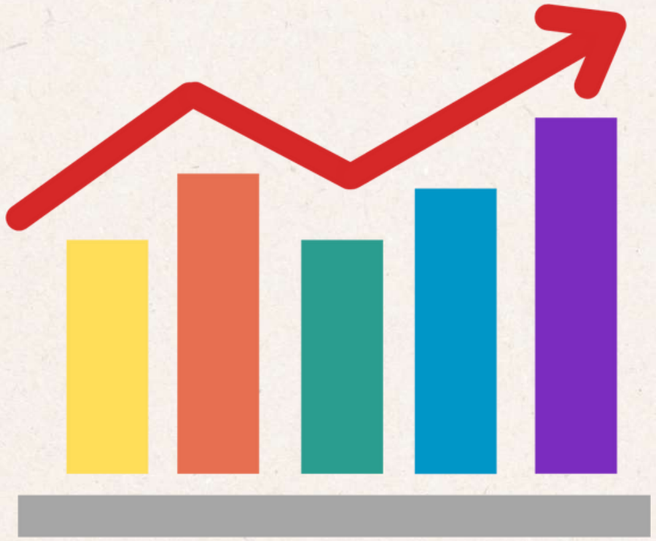
CROSS-INDUSTRY STANDARD PROCESS FOR DATA MINING





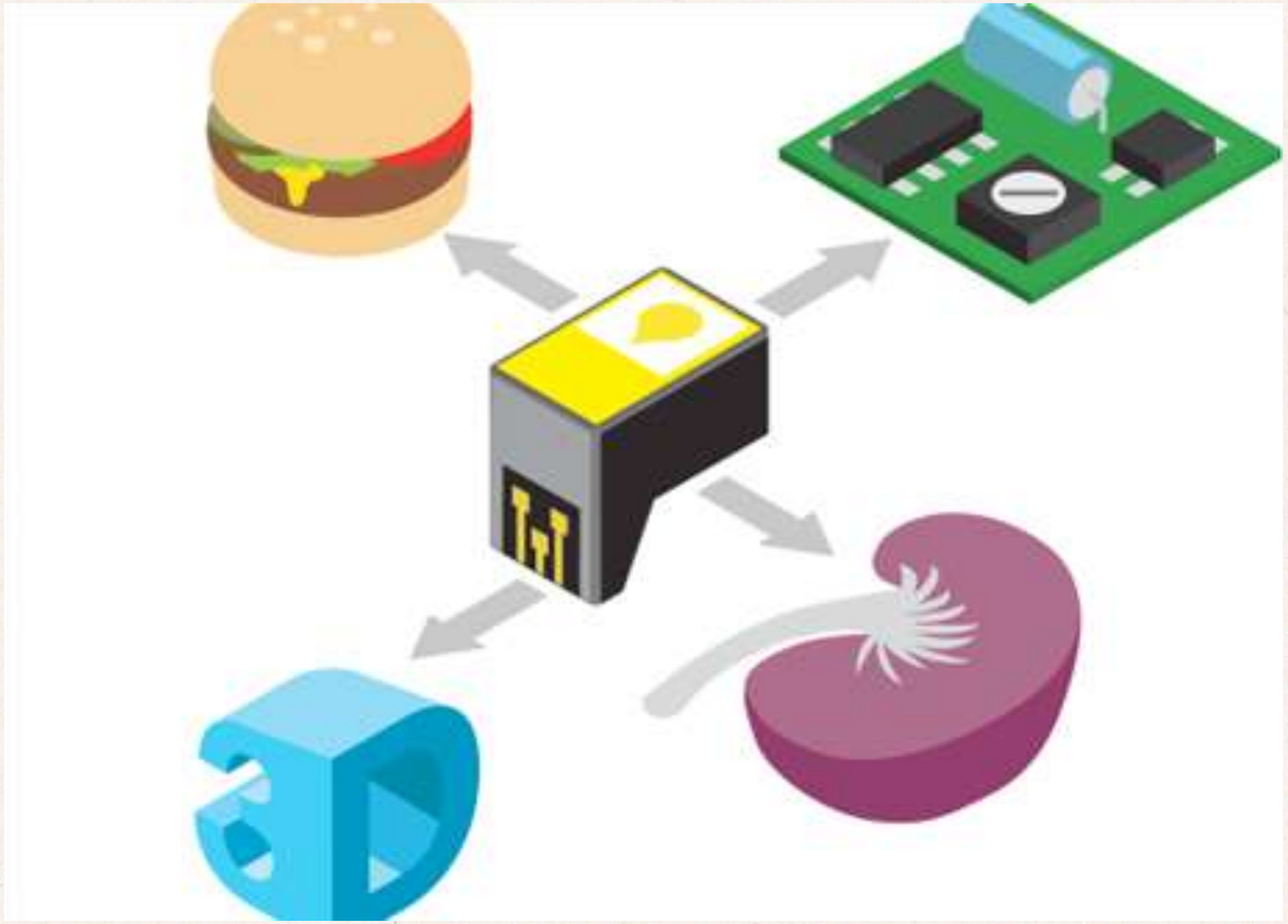
What is Data Mining?

The process of discovering patterns, trends, or relationships hidden inside large datasets.

Data mining = finding valuable knowledge inside mountains of information.

Cross Industry

Not just for one sector

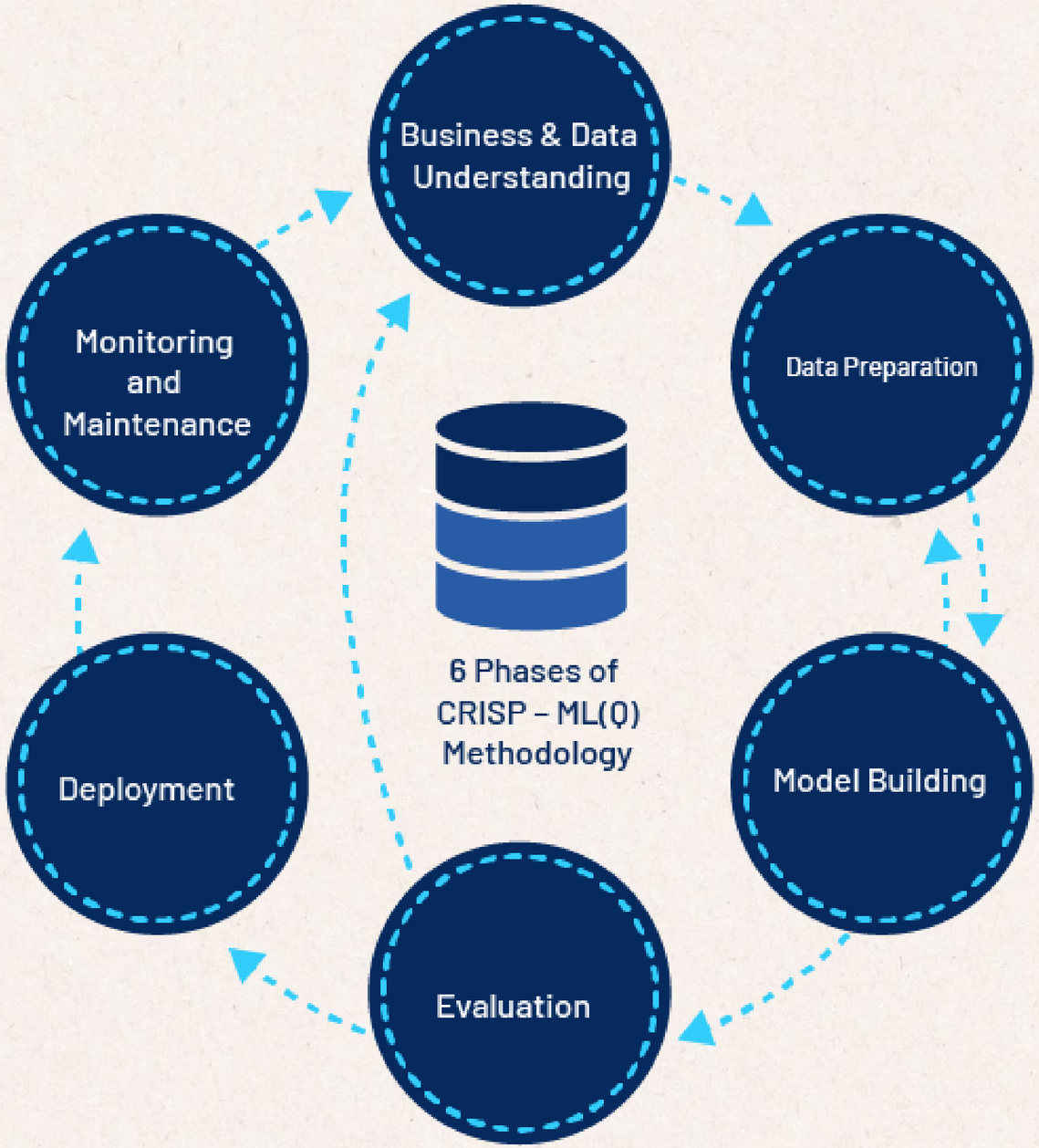


Standard Process

Alignment with business goals

A structured workflow

Consistency and quality

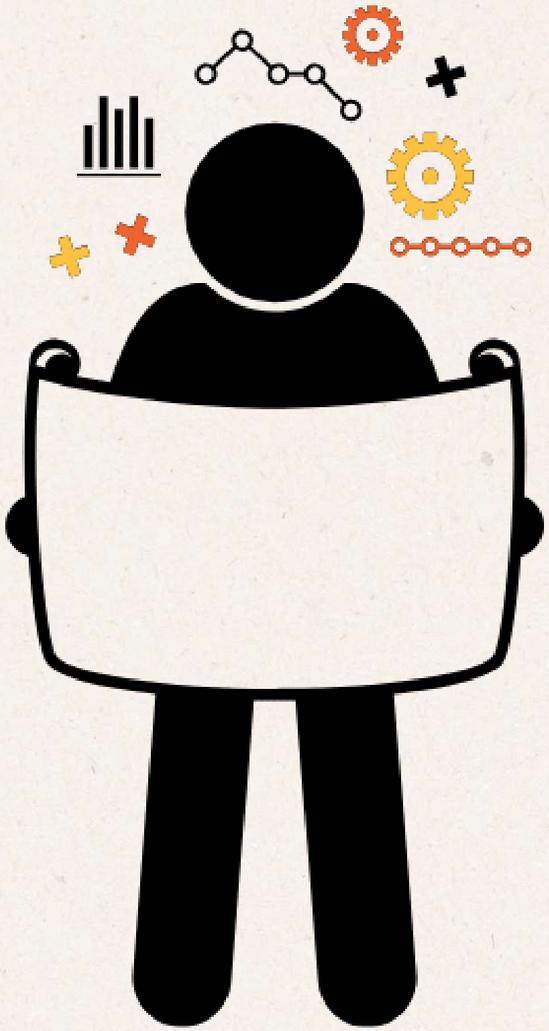


Business Understanding

What problem are we solving?
Why does it matter?
How will we know if we've succeeded?

Data Understanding

In this phase, analysts collect the initial data and then spend time exploring it – checking its structure, quality, and completeness.



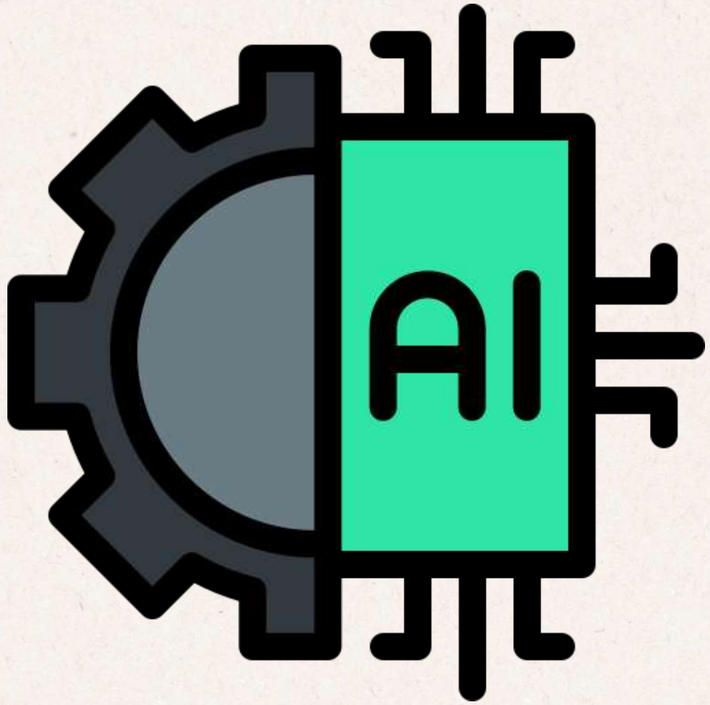
Data Preparation

- Cleaning – fixing missing values, removing duplicates, and correcting errors.
- Transforming – converting data into usable formats, such as normalizing numeric values or encoding categories.
- Integrating – merging multiple data sources so they align correctly.
- Feature Selection or Creation – choosing the most relevant variables, or engineering new ones that may better explain the problem.

Modeling

In this phase, analysts select one or more algorithms and train them using the cleaned and structured data.

The goal is to build a model that captures the relationships between the input variables (like customer behavior, usage frequency, or complaint patterns) and the target variable

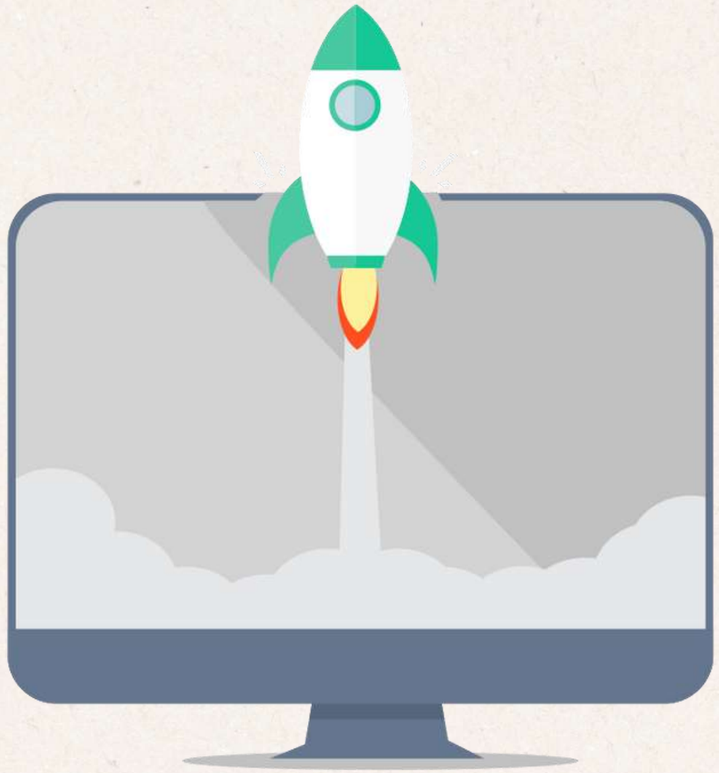


Evaluation

We assess our chosen model not just by how accurate it is, but by how well it aligns with the business objectives we established during the Business Understanding phase.

Deployment

Delivering the model’s insights or outputs in a way that helps the business.



“So to wrap everything up, the CRISP-DM framework gives us a complete, structured roadmap for turning raw data into real business value. It’s not just a technical process, it’s a disciplined way of thinking about how to connect data, insight, and decision-making.”

Horizon College has noticed an increase in student dropouts over the past two years. The administration wants to use data analytics to identify which students are most at risk of dropping out so they can provide early academic and emotional support.

Business Understanding

The main goal is to predict which students are most likely to drop out before the semester ends. Success will be measured by the college's ability to reduce dropout rates by at least 10% in the next academic year.
If the model helps academic advisers contact at-risk students early and improves retention, the project will be considered successful.

Data Understanding

The college will gather historical student data, including attendance records, GPA per semester, financial aid status, counseling visits, extracurricular participation, and number of failed subjects.
Early exploration might reveal that poor attendance and low GPA are common among students who eventually drop out.
There may also be missing records or inconsistent data between departments.

Data Preparation

Data will be cleaned by removing duplicate entries and filling in missing attendance or grade information.

Irrelevant columns like student address or ID number will be excluded.

New variables will be created, such as “average grade trend” and “absences per month.”

Data from different departments (registrar, guidance, finance) will be combined into one organized table.

Modeling

The cleaned data will be fed into a computer system to help discover patterns.

Here are the key questions we would ask the machine:

1. Do students with declining GPA across semesters have higher dropout probability?
2. How strongly does attendance rate affect dropout risk?
3. Does receiving financial aid reduce the likelihood of dropping out?
4. Are students who don’t join extracurricular activities more likely to leave school?
5. Can the model predict which students might drop out before midterms?

These questions will help the team identify early warning signs that can guide interventions.

Evaluation

To check if the model works, we will test its predictions using unseen data from recent semesters.
If the model correctly identifies at least 80% of students who eventually dropped out, it will be considered effective.
We will also evaluate whether the predictions make sense from an academic point of view, not just statistically.
If the model is accurate but flags students unfairly or ignores social factors, it may need refinement.

Deployment

Once validated, the model's results will be shared with the Guidance Office and Academic Affairs team.

Each semester, the college can generate a list of high-risk students and assign mentors to contact them early.

The model will be updated with new data at the end of every term to keep it accurate.

The project's long-term goal is to integrate the model into the student information system so risk alerts appear automatically for advisers.