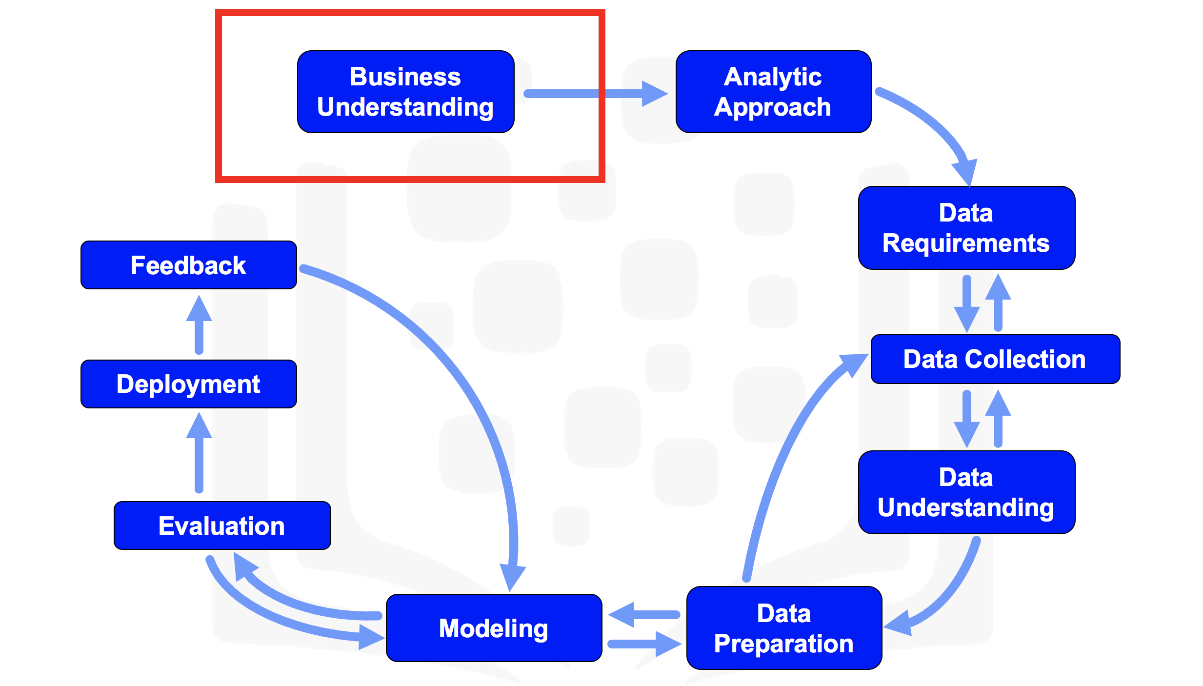
DATA SCIENCE – IBM

Data Science methodology begins with spending time to seek clarifications to attain a business understanding.



**STEPS:**

**1) BUSINESS UNDERSTANDING**

The first step in the data science methodology is achieving a clear understanding of the business problem. This involves identifying the actual goal behind a question.

To demonstrate this, consider a healthcare case study. An American insurance provider faced funding cuts and wanted to optimize their limited healthcare budget without raising customer premiums. After discussions, the team identified patient readmissions, particularly among Congestive Heart Failure (CHF) patients, as a key area to address.

A decision-tree model was selected to analyse the causes and predict readmission risks

Four core business requirements were defined:

1. Predict readmission outcomes for CHF patients
2. Assess individual readmission risk
3. Understand the contributing factors behind the predictions
4. Ensure the model is easy to apply to new patient data

**2) ANALYTIC APPROACH**

The second stage of the data science methodology is selecting the right analytic approach, which depends on the type of question being asked.

* If the goal is to predict an outcome → use a predictive model
* If the aim is to understand relationships → use a descriptive model
* For yes/no answers → use a classification model
* To explore human behaviour patterns → use clustering or association techniques

Machine Learning plays a vital role here, as it helps uncover hidden patterns and trends in data without explicit programming.

Case Study Example:

To predict readmission risk for Congestive Heart Failure (CHF) patients, a decision tree classification model was used.

* The model identifies conditions leading to readmission and provides both:
  + The predicted outcome
  + The likelihood (risk) of readmission
* It's easy for non-data scientists (like clinicians) to interpret and apply
* Multiple models can be used at different stages to track how a patient's risk evolves during treatment

**3)DATA REQUIREMENTS**

Focuses on identifying the right data needed to solve the problem. Once the business problem and analytic approach are established, the data scientist determines:

* What data is needed
* Where it will come from (source)
* In what format it should be collected and prepared

Case Study Example:

1. Patient Selection Criteria:
   * Admitted as in-patient within the provider’s service area
   * Primary CHF diagnosis within one year
   * At least six months of continuous prior enrollment for full clinical history
2. Exclusions:
   * Patients with other major medical conditions were excluded to avoid skewed readmission rates.
3. Data Needed:
   * One record per patient
   * Columns for diagnoses, procedures, prescriptions, and hospital visits
   * Thousands of transactional records were aggregated into a single row per patient by creating new variables as part of the data preparation stage

**4) DATA COLLECTION**

The actual data needed for analysis.

Key Actions in This Stage:

* Assess the availability and quality of the collected data
* Use descriptive statistics and visualization to explore content and identify gaps
* Decide whether more data is needed or if missing parts can be filled later

Case Study Example:

* Data was collected from multiple sources:
  + Demographics, clinical records, claims, provider info, and pharmaceutical data
* Some drug data was initially unavailable, but the model was still built successfully
* If needed, deferred data (like drug info) can be added later based on model results

LINK TO LAB : [ai-agents-/Data Science/DS0103EN-2 0Review-From-Requirements-to-Collection-R.ipynb at main · ibilji/ai-agents-](https://github.com/ibilji/ai-agents-/blob/main/Data%20Science/DS0103EN-2%200Review-From-Requirements-to-Collection-R.ipynb)

**5) DATA UNDERSTANDING**

* Includes all activities related to constructing the dataset.
* It answers the question: "Is the data collected representative of the problem to be solved?"

Case Study Example:

To understand the dataset:

1. Descriptive statistics were applied to variables:
   * Measures like mean, median, min, max, standard deviation.
   * Used to understand univariate properties and detect issues.
2. Pairwise correlations:
   * Identify variables that are highly correlated, indicating redundancy.
   * Redundant variables are often dropped to simplify the model.
3. Histograms:
   * Visualize distribution of variable values.
   * Help assess data quality and necessary preparation steps.
   * E.g., Categorical variables with too many unique values might be grouped or consolidated.

Data Quality Considerations:

* Assess for missing values and invalid entries.
* Decide whether missing means:
  + A valid "no", "0", or
  + Truly unknown.

Iterative Process:

* Initially, congestive heart failure admission was defined by primary diagnosis.
* Upon analysis, the definition was revised to include secondary and tertiary diagnoses.
* This looped back to the data collection stage, showing how understanding evolves over time.

**6) DATA PREPARATION**

This stage is time-intensive as it often takes 70% to 90% of the entire project time.

Important activities during data preparation include:

* Handling missing or invalid values
* Removing duplicates
* Ensuring consistent formatting
* Feature engineering, which is especially vital when working with machine learning models

When working with text data, text analysis and proper coding are also necessary to extract meaningful features and avoid overlooking hidden insights.

LINK TO LAB : [ai-agents-/Data Science/DS0103EN-Exercise-From-Understanding-to-Preparation (copy) (copy).ipynb at main · ibilji/ai-agents-](https://github.com/ibilji/ai-agents-/blob/main/Data%20Science/DS0103EN-Exercise-From-Understanding-to-Preparation%20(copy)%20(copy).ipynb)

**7) MODELLING**

Modeling is the phase in the data science methodology where a data scientist builds and tests models to extract insights or make predictions.

* Purpose of Modeling:  
  To create descriptive models (e.g., behavior patterns) and predictive models (e.g., yes/no outcomes).
* Approaches:
  + Statistical methods or machine learning algorithms are used.
  + Training datasets (with known outcomes) are critical for building and evaluating predictive models.
* Model Building Involves:
  + Trying different algorithms.
  + Selecting relevant variables.
  + Constant tuning and refinement.

**8) EVALUATION**

Model evaluation is an essential step in the data science process, closely tied with model building. These two stages are typically carried out iteratively , they are repeated to refine the model and improve its performance.

Purpose of Evaluation

The goal of evaluation is to assess the quality and effectiveness of a model before it's deployed. It answers critical questions such as:

* Does the model meet the original objectives?
* Does it provide the right answers, or does it require adjustment?

Phases of Model Evaluation

Evaluation generally involves two main phases:

1. Diagnostic Measures:
   * These measures ensure the model functions as expected.
   * For predictive models, techniques like decision trees can help assess whether the model's outputs align with the initial design and identify areas needing improvement.
   * For descriptive models, testing sets with known outcomes are used to evaluate and fine-tune the model.
2. Statistical Significance Testing:
   * This ensures that the data is correctly handled and interpreted.
   * It reduces doubts and unnecessary second-guessing when interpreting model outcomes.

ROC Curve: Selecting the Optimal Model

* The Receiver Operating Characteristic (ROC) curve , used to evaluate performance.
* It plots true-positive rate vs. false-positive rate as the decision threshold changes.
* The optimal model is the one where the ROC curve shows the maximum separation from the baseline .

LINK TO LAB : [ai-agents-/Data Science/DS0103EN-Exercise-From-Modeling-to-Evaluation.ipynb at main · ibilji/ai-agents-](https://github.com/ibilji/ai-agents-/blob/main/Data%20Science/DS0103EN-Exercise-From-Modeling-to-Evaluation.ipynb)

**9) DEPLOYMENT**

The model may first be deployed in a limited test environment before full rollout to ensure confidence in its results. Ensure the model is usable, interpretable, and actionable for the intended users, leading to measurable impact (e.g., reduced readmissions).

**10) FEEDBACK**

Once a model is deployed, feedback from users becomes essential for refining it and evaluating its performance and impact. This continuous feedback loop ensures that the model remains effective and relevant over time.