🧠 Data Science Methodology – Module 1 Notes

🌍 1. Introduction to Data Science Methodology

📘 Purpose of the Course

- Understand how **data scientists use a structured approach** to solve business problems.
- Learn **10 stages** of a standard data science methodology.
- Apply concepts using a **real-life healthcare case study** (allocating limited budget to improve patient care).

🧩 2. What is a Methodology?

Definition

- **Methodology**: A system of methods used to guide scientific or research work.
- In data science, it provides a **structured framework** for solving complex problems **using data**.

Importance

- Prevents jumping directly to solutions without understanding the problem.
- Ensures **clarity**, **consistency**, and **valid results**.
- Guides every stage from **problem definition to feedback**.

🧱 3. The 10 Stages of Data Science Methodology (John Rollins – IBM)

Stage	Question it answers	Purpose
1️⃣ Business Understanding	What is the problem?	Define business goals clearly.
2️⃣ Analytic Approach	How can we use data to solve it?	Choose the right analytical path.
3️⃣ Data Requirements	What data do we need?	Identify necessary data types, formats, and sources.
4️⃣ Data Collection	Where and how do we get data?	Gather data and verify completeness.
5️⃣ Data Understanding	Does data represent the problem well?	Explore data for patterns and issues.
6️⃣ Data Preparation	What work is needed to make data usable?	Clean, format, and combine data.
7️⃣ Modeling	Can we build a model to answer the question?	Apply suitable modeling techniques.
8️⃣ Evaluation	Does the model solve the problem?	Assess accuracy and business usefulness.
9️⃣ Deployment	Can we put it into practice?	Implement the model in real-world use.
🔟 Feedback	What improvements can be made?	Refine based on new data and outcomes.

💼 4. Stage 1: Business Understanding

Purpose

- Clarify **the real business problem** before using data.
- Ensure everyone agrees on **the goal and objectives**.

Process

1. **Define the goal** – What is the main issue or aim?
2. **Identify objectives** – What steps support this goal?
3. **Engage stakeholders** – Clarify requirements and constraints.

Common Mistake

- Solving the **wrong problem** because of poor question clarity.

🏥 Case Study Example

- Problem: *How to best allocate a limited healthcare budget to improve patient care?*
- Stakeholders: Insurance providers, healthcare authorities, IBM data scientists.
- Key Requirements for the Model:
    1. Predict readmission outcomes (Congestive Heart Failure patients).
    2. Predict readmission risk.
    3. Identify combinations of factors leading to readmission.
    4. Provide an easy-to-use process for future patients.

⚙️ 5. Stage 2: Analytic Approach

Purpose

- Select **how** to analyze data depending on the **type of question**.

Types of Analytical Approaches

Question Type	Approach	Techniques	Example

Descriptive (“What happened?”)	Descriptive Analytics	Aggregation, Visualization	Monthly sales report
Diagnostic (“Why did it happen?”)	Diagnostic Analytics	Correlation, Drill-down	Cause of sales drop
Predictive (“What will happen?”)	Predictive Analytics	Regression, Forecasting, ML	Sales prediction
Prescriptive (“What should we do?”)	Prescriptive Analytics	Optimization, Simulation	Pricing strategy
Classification (“Which category?”)	Classification	Decision Trees, SVMs, Neural Nets	Spam detection

🧮 Case Study Example

- Chosen approach: **Decision Tree Classification Model**
- Why?
    - Easy to interpret.
    - Predicts readmission likelihood (yes/no).
    - Shows factors contributing to outcomes.
    - Clinicians can understand and use it easily.

📊 6. Stage 3: Data Requirements

Purpose

- Identify what **data is needed**, its **format**, and **where it comes from**.
- Think of this step like **gathering ingredients before cooking**.

Case Study Example

- Needed patient data for **decision tree classification**.
- Criteria for patient cohort:
    1. In-patient within the provider's service area.
    2. Primary diagnosis: congestive heart failure.
    3. Continuous 6-month enrollment before admission.
- Excluded patients with multiple serious diseases (to avoid bias).
- Required **one record per patient** summarizing medical history.

📥 7. Stage 4: Data Collection

Purpose

- Gather the identified data from defined sources.
- Check **data completeness and quality**.

Activities

- Extract data from multiple sources (e.g., demographic, clinical, claims, pharmaceutical).
- Merge data and remove redundancy.
- Identify missing or unavailable data.
- Decide whether missing data is critical or can be deferred.

Case Study Note

- Missing **drug data** initially — team continued without it.
- Could add it later if model performance suggested it mattered.

📘 Data Science Methodology – Module 2 Notes

🧩 From Understanding to Preparation → Modeling to Evaluation

🧠 1. Data Understanding

📌 Definition

Data Understanding involves **all activities related to examining and exploring data** to determine:

> "Is the data collected representative of the problem we're trying to solve?"

It helps verify the **accuracy, completeness, and relevance** of the data before modeling.

🧮 Key Activities in Data Understanding

Step	Description	Purpose
1. Descriptive Statistics	Compute basic statistics (mean, median, min, max, std. deviation)	Understand data characteristics
2. Correlation Analysis	Check how variables are related	Detect redundant variables

3. Histograms	Visualize data distribution	Identify skewness, outliers, or too many unique values
4. Data Quality Checks	Detect missing, invalid, or outlier values	Decide whether to clean, recode, or remove data

---

📊 Example — Case Study: Congestive Heart Failure (CHF)

- Descriptive statistics and correlations were run on patient data.
- Histograms helped identify how variables were distributed.
- Some categorical variables had **too many distinct values** — histograms helped determine how to **combine or simplify them**.
- Detected **invalid values**, such as `age = 999`, meaning *missing* rather than a true value.
- Adjusted definition of **CHF admissions** to include **secondary and tertiary diagnoses**, improving data completeness.

🔄 Key Insight

Data Understanding is **iterative** — it often requires looping back to **Data Collection** to refine definitions or include missing information.

🗒️ Summary

✅ Data Understanding ensures:

- The dataset truly represents the problem.
- Unnecessary, redundant, or invalid data is removed.

- The model will be built on **clean, reliable, and relevant** information.

🧰 2. Data Preparation

📘 Definition

Data Preparation involves **cleaning, transforming, and organizing data** to make it ready for modeling.

> "Just as vegetables must be washed before cooking, data must be cleaned before modeling."

⏱️ Time & Effort

- Most time-consuming stage: **70% to 90%** of total project effort.
- Automation can reduce it to **50%**.

⚙️ Key Tasks in Data Preparation

Step	Description	Goal
1. Handle Missing/Invalid Values	Replace, recode, or remove nulls, zeros, or invalid data (like "999")	Improve data quality
2. Remove Duplicates	Ensure each observation is unique	Avoid bias in model
3. Format Data	Standardize data types, scales, and formats	Ensure consistency

4. Feature Engineering	Create new features based on domain knowledge	Make model more effective
5. Text Analysis (if applicable)	Process and code textual data	Prepare text for modeling

---

🧠 Feature Engineering

- A **feature** is a measurable property or characteristic that helps solve the problem.
- Example: Instead of raw hospital visit data, use features like:
    - *Number of visits per month*
    - *Time since last admission*
    - *Presence of co-morbidities (e.g., diabetes, hypertension)*

Good features = better models.

🍲 Analogy

Like chopping onions before cooking — smaller, well-prepared pieces spread flavor evenly. Likewise, properly prepared data spreads information evenly through the model.

⚕️ Case Study: Data Preparation in CHF

Key Steps:

1. **Define CHF precisely:**
    - Identify correct diagnosis codes for congestive heart failure (not just “heart failure”).

    - Consult clinical experts.

2. **Define Re-admission Criteria:**
    - A re-admission within **30 days** of discharge counts as a CHF readmission.
3. **Aggregate Transactional Records:**
    - Multiple patient records → merged into **one record per patient**.
    - Combined data from:
        - Hospital claims
        - Lab tests
        - Prescriptions
        - Diagnoses
4. **Feature Creation:**
    - Frequency of visits, co-morbidities, demographics (age, gender, insurance type).
    - Added missing conditions after a **literature review**.
5. **Final Dataset:**
    - One record per patient (2,343 patients).
    - Dependent variable: *Readmission within 30 days (Yes/No)*.
    - Split into **Training Set** and **Testing Set**.

🗒️ Summary

✅ Data Preparation ensures:

- Clean, complete, and well-structured data.
- Useful features that improve model performance.

- Model inputs match the analytic approach.

🧮 3. Modeling

📘 Definition

Modeling is the stage where **mathematical or machine learning models** are developed to describe or predict outcomes.

> “This is where the data scientist samples the sauce to see if it’s well-seasoned.”

🎯 Purpose

To create a **descriptive** or **predictive** model that answers the business question defined earlier.

Model Type	Goal	Example
Descriptive	Find patterns or relationships	“If a person does this, they are likely to prefer that.”
Predictive	Predict future outcomes	Predict whether a patient will be re-admitted.

⚙️ Key Elements

- Uses **training data** with known outcomes.
- Tests different **algorithms** (e.g., Decision Trees, Regression, SVM).
- Involves **parameter tuning** — adjusting settings to improve performance.
- Requires **iteration** with earlier stages (data prep, feature selection).

🧠 Analogy

Just like adjusting seasoning in food until it tastes right — modeling involves **refining and adjusting** until the model performs optimally.

⚕️ Case Study: CHF Model Building

Step 1 — Initial Model:

- Used **Decision Tree Classification**.
- Accuracy = **85%**, but only **45% sensitivity** (correctly predicting "yes" readmissions).
  → The model wasn't accurate enough.

Step 2 — Adjusting Cost Parameters:

- Adjusted **misclassification costs**:
  - False positives (predicting yes when it's no)
  - False negatives (predicting no when it's yes)
- Tried different ratios:
  1️⃣ **9:1** → 97% "yes" correct, but only 49% overall accuracy.
  2️⃣ **4:1** → Balanced: 68% sensitivity ("yes"), 85% specificity ("no"), 81% overall.

✅ Final choice: **4:1 ratio** — best trade-off between accuracy and cost.

⚖️ Type I vs Type II Errors

Type	Description	Example
Type I (False Positive)	Predicts a patient *will* be readmitted, but they won't.	Wasted intervention

Type II (False Negative)	Predicts a patient *won't* be readmitted, but they are.	Missed treatment → costly error

---

📈 4. Evaluation

📘 Definition

Evaluation determines whether the **model meets business objectives** and performs well on unseen data.

> "Does the model truly answer the original question?"

🔍 Phases of Evaluation

Phase	Purpose	Example
1. Diagnostic Measures	Check if the model works as intended	Accuracy, sensitivity, specificity
2. Statistical Significance Testing	Ensure reliability and validity	p-values, confidence intervals

---

🧠 Tools and Metrics

- **ROC Curve (Receiver Operating Characteristic):**
  - Plots **True Positive Rate** vs **False Positive Rate**.
  - The **higher and further left** the curve, the better the model.

- The **best model** gives **maximum separation** from the baseline.

⚕️ Case Study: CHF Model Evaluation

- Compared models with different misclassification cost ratios.
- Used **ROC curves** to find the **optimal model (4:1 ratio)**.
- Confirmed model was both **accurate** and **cost-effective** for the hospital's budget.

🗒️ Summary

✅ Evaluation ensures:

- The model meets original goals.
- The model generalizes well to unseen data.
- The right balance is achieved between sensitivity and specificity.

🧩 Overall Module Summary

Stage	Key Question	Outcome
Data Understanding	Is the data representative of the problem?	Insights on quality and completeness
Data Preparation	How do we clean and transform the data?	Usable, structured dataset

Modeling	Can we build a model to solve the problem?	Predictive or descriptive model
Evaluation	Does the model meet objectives?	Validated and optimized model

📘 Data Science Methodology – Module 3 Notes (Deployment → Feedback → Storytelling → Summary)

🚀 1. Deployment Stage

🎯 Purpose

- To **put the model into real-world use** and make it meaningful for stakeholders.
- Goal: **Translate model results into practical action.**

🔑 Key Ideas

- Stakeholders (solution owners, IT, developers, clinicians, etc.) must understand how to use the model.
- Deployment may begin in a **limited or test environment** before full rollout.
- Ensures that **business users** can confidently apply model insights in decision-making.

🏥 Case Study – Congestive Heart Failure (CHF)

- Model deployment aimed to **reduce 30-day readmission rates**.

- The business team translated model outputs for clinical staff:
    - Helped identify **high-risk patients**.
    - Guided design of **intervention programs**.
- Requirements for the deployment app:
    - **Automated, near real-time** risk assessment.
    - **Tablet-based browser app** for convenience.
    - Data automatically formatted and scored **before patient discharge**.
- Training and tracking processes were created with **IT support** and **database teams** to monitor performance and outcomes.

💡 Example: Cognos Application

- Another model (for juvenile diabetes) used **decision tree classification** to predict hospitalization risk.
- Visual maps and summary reports helped clinicians:
    - View **national risk patterns**.
    - Analyze **specific subgroups** of patients.
    - Review **individual patient summaries**.

🔁 2. Feedback Stage

🎯 Purpose

- To **evaluate real-world performance** and **refine the model** based on actual outcomes.
- Makes the methodology **cyclical and iterative**.

🔑 Key Points

- Continuous feedback ensures **ongoing model improvement**.
- The process includes **monitoring**, **measurement**, and **adjustment**.

🏥 Case Study Steps

1. **Define a review process** — executives monitor results of the CHF risk model.
2. **Track intervention outcomes** — record re-admission rates of patients receiving interventions.
3. **Measure effectiveness** — compare readmission rates **before vs. after** implementation (no control group due to ethics).
4. **Refine the model** — based on:
    - New data from interventions.
    - Possibly add **pharmaceutical data** that was initially skipped.
    - Adjust intervention actions and procedures.
5. **Redeploy improved model** — feedback continues through the program's life.

🔄 Insight

> Feedback is continuous — as knowledge grows, the model and actions evolve.

🗣️ 3. Storytelling in Data Science

🎯 Why It Matters

- **Storytelling turns data into understanding.**
- Helps stakeholders connect emotionally and logically to insights.

💬 Expert Insights

- Humans process and remember **stories better than statistics**.
- Effective storytelling:
    - Makes complex data **clear and persuasive**.
    - Bridges gap between **technical results** and **business impact**.
- A story gives **context**, **emotion**, and **meaning** to numbers.

🧩 Core Skills

- Be **clear, concise, and compelling**.
- Maintain **balance** — simplify enough to engage, but not to distort facts.
- Use **visuals and examples** to communicate insights effectively.

💡 Real Lesson

> Anyone can show data — only a storyteller can inspire action.

🧭 4. Course Summary – Key Takeaways

💼 What You Learned

- How to **think like a data scientist**:
    1. Define the **problem** clearly.
    2. Choose the **analytic approach** wisely.
    3. Identify and **collect the right data**.
    4. **Prepare, model, evaluate, and deploy**.
    5. Gather **feedback** to improve continuously.
- Each stage is **iterative** — improvement never stops.

🧩 Methodology Essence

> 10 Questions → 10 Steps → A Complete Cycle of Data Science Work.

💬 Core Philosophy (John Rollins)

> “Your success depends on applying the right tools, at the right time, in the right order, to the right problem.”