CBSE AI 843 Sample Paper). These are **not Q&A**, but **proper notes for study/project use**

High Priority Topics (Must Prepare 100%)

 1. Data Science Methodology (VERY IMPORTANT)
 Focus:
- All **10 steps** (order + meaning)
- Difference between:
  - Business Understanding vs Analytic Approach
- Types of Analytics:
  - Descriptive
  - Diagnostic
  - Predictive
  - Prescriptive

 क्यों महत्वपूर्ण:
 हर साल questions आते हैं (MCQ + 2 marks + 4 marks)

 2. Making Machines See (Computer Vision)
 Focus:
- Pixel concept
- Image processing techniques:
  - Edge Detection
  - Corner Detection
  - Texture Analysis
  - Colour Extraction
- Tasks:
  - Image Classification
  - Object Detection

 क्यों:
 Diagram + MCQ based questions आते हैं

 3. Big Data & Analytics
 Focus:
- Definition of Big Data
- Sources:
  - Social data
  - Machine data
  - Transactional data
- Moore's Law
- Characteristics (Volume, Variety, Velocity, Veracity)

 क्यों:

 Case-based questions आते हैं

 4. Neural Networks (CORE TOPIC)
 Focus:
- ANN vs Deep Learning
- Hidden layers
- Activation Functions:

    - Sigmoid
    - Tanh
    - ReLU
- Learning rule

 क्यों:
 Conceptual + MCQ + short answer

 5. Generative AI (TRENDING & SCORING)
 Focus:
- What is Generative AI
- GAN:
    - Generator
    - Discriminator
- Applications:
    - Text
    - Image
    - Video
    - Audio

 क्यों:
 New topic = high chances in exam

 6. Data Storytelling
 Focus:
- Components:
    - Data
    - Visual
    - Narrative
- Term: **Insight**
- Freytag's Pyramid (very important for 4 marks)

 क्यों:
 Long answer question आता है

⚡ **Medium Priority Topics**
 Model Validation
- Train-Test Split
- K-Fold Cross Validation
- Differences (very important)

 Evaluation Metrics
 Classification:
- Confusion Matrix
- Accuracy
- Precision
- Recall
- F1 Score

 Regression:
- MSE
- RMSE

 Feature Engineering

- Definition + examples

 Employability Skills (Easy Marks – Don't Ignore)
 Focus:
- Communication (Active listening – RESPECT)
- Self-management:
  - Positive thinking
  - Self-awareness
- ICT:
  - Slide layout
- Entrepreneur:
  - Traits
- Green Skills:
  - Appropriate technology

 यहाँ से direct MCQ आते हैं

 DATA SCIENCE METHODOLOGY – COMPLETE NOTES
 Objective of Data Science Methodology
The objective of Data Science Methodology is to:
- Provide a **structured and systematic approach** to solve real-world problems using data
- Help in **understanding the problem clearly before solving it**
- Guide through all stages:
  - Data collection
  - Data preparation
  - Model building
  - Evaluation
- Ensure **accurate, reliable, and meaningful results**
- Support **data-driven decision making**

 In short:
"To convert raw data into useful insights and solutions using a step-by-step process."

डेटा साइंस मेथडोलॉजी का उद्देश्य है:
- डेटा की मदद से समस्याओं को हल करने के लिए **एक व्यवस्थित प्रक्रिया प्रदान करना**
- समस्या को सही तरीके से समझना
- सभी चरणों में मार्गदर्शन देना:
  - डेटा संग्रह
  - डेटा तैयारी
  - मॉडल बनाना
  - मूल्यांकन
- **सही और भरोसेमंद परिणाम प्राप्त करना**
- डेटा के आधार पर सही निर्णय लेना

 सरल शब्दों में:
"कच्चे डेटा को उपयोगी जानकारी और समाधान में बदलना।"

 Exam Tip (Important Line)

 If 2 marks question comes, write this line:
English:
"Data Science Methodology aims to provide a structured approach to solve problems using data and generate meaningful insights."
Hindi:
"डेटा साइंस मेथडोलॉजी का उद्देश्य डेटा का उपयोग करके समस्याओं को हल करना और उपयोगी जानकारी प्राप्त करना है।"

 1. What is Data Science Methodology?

 English:
Data Science Methodology is a **structured, step-by-step process** used by data scientists to solve problems using data.
It ensures the project is **systematic, efficient, and goal-oriented**.
 हिंदी:
डेटा साइंस मेथडोलॉजी एक **क्रमबद्ध प्रक्रिया** है जिसका उपयोग डेटा की सहायता से समस्याओं को हल करने के लिए किया जाता है।
यह प्रोजेक्ट को **सही दिशा और व्यवस्थित तरीके से पूरा करने में मदद करती है।**

 2. All 10 Steps (Order + Meaning) IMPORTANT

 1. Business Understanding
English:
- Identify the **problem to be solved**
- Understand goals and requirements

हिंदी:
- समस्या को समझना
- लक्ष्य और आवश्यकता पहचानना

 2. Analytic Approach
English:
- Decide **how to solve the problem using data**
- Choose approach: classification, regression, clustering

हिंदी:
- समस्या हल करने का तरीका चुनना
- जैसे: classification, regression

 3. Data Requirements
English:
- Identify:
  - Type of data (text, image, numbers)
  - Format and source

हिंदी:
- किस प्रकार का डेटा चाहिए
- उसका स्रोत और प्रारूप

🔹 4. Data Collection
English:
- Gather data from:
  - Primary sources (survey, sensors)
  - Secondary sources (internet, database)

हिंदी:
- डेटा इकट्ठा करना
- स्रोत:
  - प्राथमिक (survey)
  - द्वितीयक (internet)

🔹 5. Data Understanding
English:
- Check:
  - Data quality
  - Missing/incorrect values
- Use charts, statistics

हिंदी:
- डेटा की गुणवत्ता जांचना
- गलत या missing values देखना

🔹 6. Data Preparation
English:
- Clean data:
  - Remove errors
  - Fill missing values
- Perform **Feature Engineering**

हिंदी:
- डेटा साफ करना
- नई विशेषताएँ बनाना (Feature Engineering)

🔹 7. Modeling
English:
- Build model using algorithms
- Use training data

हिंदी:
- मॉडल बनाना
- एल्गोरिद्म का उपयोग

🔹 8. Evaluation
English:
- Check model performance:
  - Accuracy
  - Precision
- Improve if needed

हिंदी:
- मॉडल का मूल्यांकन
- सुधार करना

 9. Deployment

English:

- Apply model in real-world system

हिंदी:

- मॉडल को वास्तविक उपयोग में लाना

 10. Feedback

English:

- Collect user feedback
- Improve model continuously

हिंदी:

- उपयोगकर्ता प्रतिक्रिया लेना
- मॉडल को सुधारना

3. Difference: Business Understanding vs Analytic Approach VERY IMPORTANT

Feature	Business Understanding	Analytic Approach
English	Defines **problem**	Defines **solution method**
Focus	What to solve	How to solve
Example	Predict student marks	Use regression
Stage	First step	Second step

 हिंदी:

बिंदु	Business Understanding	Analytic Approach
क्या	समस्या क्या है	समाधान कैसे होगा
फोकस	Problem	Method
उदाहरण	छात्र के अंक का अनुमान	Regression
चरण	पहला	दूसरा

 4. Types of Analytics VERY IMPORTANT

 1. Descriptive Analytics

English:

- Answers: **What happened?**
- Uses past data
- Example: Average marks

हिंदी:

- क्या हुआ?
- पिछले डेटा का विश्लेषण

 2. Diagnostic Analytics

English:

- Answers: **Why did it happen?**
- Finds causes
- Example: Why sales decreased

हिंदी:

- क्यों हुआ?
- कारण पता करना

□ 3. Predictive Analytics

English:

- Answers: **What will happen?**
- Uses ML models
- Example: Predict exam score

हिंदी:

- क्या होगा?
- भविष्य की भविष्यवाणी

□ 4. Prescriptive Analytics

English:

- Answers: **What should we do?**
- Suggests actions
- Example: Best strategy for profit

हिंदी:

- क्या करना चाहिए?
- निर्णय लेने में मदद

⚡ Quick Revision Trick

□ Steps Shortcut:

B A D C U P M E D F

(Business → Analytic → Data Req → Collection → Understanding → Preparation → Modeling → Evaluation → Deployment → Feedback)

1. Data Science Methodology

This structured approach was developed by **John B. Rollins** from IBM Analytics to turn raw data into actionable insights.

The 10 Steps (In Order):

1. **Business Understanding:** Defining the problem and goals.
2. **Analytic Approach:** Choosing the right analytical path (e.g., predictive vs. descriptive).
3. **Data Requirements:** Identifying the type and amount of data needed.
4. **Data Collection:** Gathering data from various sources.
5. **Data Understanding:** Evaluating data quality and relevance through statistics and visualization.
6. **Data Preparation:** Cleaning, transforming, and feature engineering (the most time-consuming stage).
7. **AI Modelling:** Building models like Linear Regression to make predictions.
8. **Evaluation:** Assessing model performance (e.g., using K-Fold Cross Validation).
9. **Deployment:** Integrating the model into a real-world application.
10. **Feedback:** Monitoring performance and making adjustments.

Types of Analytics:

- **Descriptive:** What happened? (e.g., summarizing historical sales) .
- **Diagnostic:** Why did it happen?
- **Predictive:** What will happen? (e.g., predicting house prices) .
- **Prescriptive:** What should we do? (e.g., prescribing efficient delivery routes) .

1. डेटा साइंस कार्यप्रणाली (Data Science Methodology)

यह डेटा वैज्ञानिकों को AI प्रोजेक्ट डिजाइन करने के लिए एक संरचित ढांचा प्रदान करता है। इसे IBM के जॉन बी. रोलिन्स (John B. Rollins) ने विकसित किया था।

इसके 10 चरण (क्रम में):

1. **Business Understanding:** समस्या और लक्ष्यों को समझना।
2. **Analytic Approach:** सही विश्लेषणात्मक रास्ता चुनना (जैसे: Predictive या Descriptive)।
3. **Data Requirements:** आवश्यक डेटा के प्रकार और मात्रा की पहचान करना।
4. **Data Collection:** विभिन्न स्रोतों से डेटा एकत्र करना।
5. **Data Understanding:** आंकड़ों और विज़ुअलाइज़ेशन के माध्यम से डेटा की गुणवत्ता की जांच करना।
6. **Data Preparation:** डेटा की सफाई और उसे मॉडलिंग के लिए तैयार करना (सबसे अधिक समय लेने वाला चरण)।
7. **AI Modelling:** भविष्यवाणियां करने के लिए एल्गोरिदम का उपयोग करके मॉडल बनाना।
8. **Evaluation:** मॉडल के प्रदर्शन का आकलन करना।
9. **Deployment:** मॉडल को वास्तविक दुनिया में लागू करना।
10. **Feedback:** परिणामों के आधार पर मॉडल में सुधार करना।

Analytics के प्रकार:

- **Descriptive (वर्णनात्मक):** क्या हुआ? (जैसे: पिछले साल की बिक्री का सारांश)।
- **Diagnostic (नैदानिक):** ऐसा क्यों हुआ?
- **Predictive (भविष्य कहनेवाला):** क्या होगा? (जैसे: घर की कीमतों की भविष्यवाणी)।
- **Prescriptive (प्रिस्क्रिप्टिव):** हमें क्या करना चाहिए? (सबसे अच्छा रास्ता बताना)।

□ 2. Making Machines See (Computer Vision)

Computer Vision (CV) enables machines to interpret visual data much like humans.

- **Pixels:** The building blocks of digital images.
- **Image Processing Techniques:**
  - **Edge Detection:** Finding boundaries of objects.
  - **Corner Detection:** Identifying specific points of interest.
  - **Texture Analysis:** Understanding the surface patterns.
  - **Colour Extraction:** Analyzing the color distribution.
- **Core Tasks:**
  - **Image Classification:** Identifying what an object is (e.g., "This is a cat").
  - **Object Detection:** Identifying what and *where* objects are in an image.

□ 2. मशीनों को देखना सिखाना (Computer Vision)

कंप्यूटर विजन मशीनों को इंसानों की तरह विज़ुअल डेटा (छवियों/वीडियो) को समझने में सक्षम बनाता है।

- **Pixels:** डिजिटल इमेज की सबसे छोटी इकाई।
- **Image Processing Techniques:**
  - **Edge Detection:** वस्तुओं की सीमाओं को खोजना।
  - **Corner Detection:** महत्वपूर्ण बिंदुओं की पहचान करना।
  - **Texture Analysis:** सतह के पैटर्न को समझना।
- **प्रमुख कार्य (Tasks):**
  - **Image Classification:** यह पहचानना कि वस्तु क्या है (जैसे: "यह एक बिल्ली है")।
  - **Object Detection:** वस्तु क्या है और वह इमेज में *कहाँ* है, दोनों को पहचानना।

□ 3. Big Data & Analytics

Big Data refers to massive, complex datasets that traditional tools cannot handle.

- **The Characteristics (6V's - Focus on 4):**
  - **Volume:** The sheer scale of data generated.
  - **Variety:** Different formats—Structured (tables), Semi-structured (emails), and Unstructured (videos) .
  - **Velocity:** The high speed at which data is generated and must be processed.
  - **Veracity:** The accuracy and trustworthiness of the data.
- **Moore's Law:** The observation that the number of transistors on a microchip doubles approximately every two years, driving the growth of data processing power.

□ 3. बिग डेटा और एनालिटिक्स (Big Data & Analytics)

बिग डेटा का मतलब ऐसे विशाल और जटिल डेटा से है जिसे सामान्य टूल्स से प्रोसेस नहीं किया जा सकता।

- **विशेषताएं (4V's):**

1. **Volume (मात्रा):** डेटा का विशाल आकार।
2. **Variety (विविधता):** विभिन्न प्रकार का डेटा (Structured, Unstructured, Semi-structured)।
3. **Velocity (वेग):** डेटा के उत्पन्न होने की उच्च गति।
4. **Veracity (सत्यता):** डेटा की सटीकता और विश्वसनीयता।

- **Moore's Law:** यह अवलोकन कि माइक्रोचिप पर ट्रांजिस्टर की संख्या लगभग हर दो साल में दोगुनी हो जाती है।

□ 4. Neural Networks

Neural Networks (NNs) are machine learning models that loosely mimic the human brain.

- **Structure:** Consists of an **Input Layer**, **Hidden Layers** (where the actual processing happens), and an **Output Layer**.
- **Components:**
  - **Weights:** Determine the signal flow between neurons.
  - **Activation Functions:** Functions like **Sigmoid**, **Tanh**, and **ReLU** determine a neuron's output.
  - **Learning Rule:** Algorithms that adjust weights based on training data.

□ 4. न्यूरल नेटवर्क (Neural Networks)

यह मानव मस्तिष्क की कार्यप्रणाली पर आधारित मॉडल हैं।

- **संरचना (Structure):** इसमें एक **Input Layer**, कई **Hidden Layers** (जहाँ प्रोसेसिंग होती है), और एक **Output Layer** होती है।
- **Activation Functions:** जैसे **Sigmoid**, **Tanh**, और **ReLU**, जो तय करते हैं कि न्यूरल नेटवर्क का आउटपुट क्या होगा।

□ 5. Generative AI

A branch of AI that creates entirely new content (text, images, audio, video) rather than just classifying existing data.

- **Generative vs. Discriminative:**
  - **Generative:** Learns to create new samples.
  - **Discriminative:** Defines boundaries to classify data (e.g., "Is this spam?").
- **GAN (Generative Adversarial Network):**
  - **Generator:** Creates fake data to try and fool the discriminator.
  - **Discriminator:** Tries to distinguish between real data and the "fake" data created by the generator.

□ 5. जनरेटिव एआई (Generative AI)

AI की वह शाखा जो नया कंटेंट (टेक्स्ट, इमेज, ऑडियो, वीडियो) बनाने में सक्षम है।

- **GAN (Generative Adversarial Network):** इसके दो भाग होते हैं:
  1. **Generator:** नया "नकली" डेटा बनाता है।
  2. **Discriminator:** असली और नकली डेटा के बीच अंतर करने की कोशिश करता है।

□ 6. Data Storytelling

The practice of translating complex data into a compelling narrative for an audience.

- **Three Key Components:**
    1. **Data:** The raw facts serving as the foundation.
    2. **Visuals:** Pictorial representations like charts and graphs.
    3. **Narrative:** The structure or "storyline" that provides context and meaning.
- **Insight:** A meaningful observation or pattern discovered in the data.
- **Freytag's Pyramid:** A 5-stage narrative structure used to organize a data story: **Exposition** (setup), **Rising Action**, **Climax**, **Falling Action**, and **Resolution** .

□ 6. डेटा स्टोरीटेलिंग (Data Storytelling)

जटिल डेटा को एक कहानी के रूप में पेश करना ताकि लोग उसे आसानी से समझ सकें।

- **3 प्रमुख घटक:** डेटा (Data), विज़ुअल्स (Visuals), और नैरेटिव/कहानी (Narrative)।
- **Freytag's Pyramid:** कहानी सुनाने की एक तकनीक जिसके 5 भाग हैं: **Exposition** (परिचय), **Rising Action**, **Climax** (चरम सीमा), **Falling Action**, और **Resolution** (समाधान)।

□ मूल्यांकन मेट्रिक्स (Evaluation Metrics)

मॉडल की सटीकता मापने के लिए उपयोग किया जाता है:

- **Confusion Matrix:** वर्गीकरण (Classification) की सटीकता देखने के लिए।
- **Accuracy, Precision, Recall:** मॉडल कितना सही है, यह जांचने के लिए।
- **MSE (Mean Squared Error):** रिग्रेशन मॉडल की गलती मापने के लिए।

Chapter 2

1. Data Science Methodology

This methodology is an iterative strategy used by data scientists to ensure that the data used in a project provides the correct answers to a problem. It was developed /pioneered by **John B. Rollins** from IBM.

The 10 Steps (In Order):

1. **Business Understanding:** The first and most critical step. It involves defining the problem, goals, and objectives from a business perspective. (e.g., "Why is customer churn increasing?")
2. **Analytic Approach:** Once the problem is defined, the data scientist identifies the best analytical path (e.g., should we use a classification model or a regression model?).
3. **Data Requirements:** Identifying the specific data needed (dates, formats, variables) to solve the defined problem.
4. **Data Collection:** Gathering the data from various sources like databases, web scraping, or surveys.
5. **Data Understanding:** Using descriptive statistics and visualization to check if the collected data actually represents the problem.
6. **Data Preparation: The most time-consuming step.** It involves cleaning the data, handling missing values, and transforming it into a format suitable for modeling.
7. **AI Modelling:** Applying algorithms (like Linear Regression or Neural Networks) to the prepared data to find patterns or make predictions.
8. **Evaluation:** Testing the model to see if it is accurate and if it meets the original business goal.
9. **Deployment:** Putting the model into action in a real-world environment (like an app or a website).
10. **Feedback:** Gathering data on how the model is performing in the real world to make further improvements.

1. डेटा साइंस कार्यप्रणाली (Data Science Methodology)

यह कार्यप्रणाली एक चक्र (cycle) की तरह काम करती है, जिसे जॉन बी. रोलिन्स ने विकसित किया था। यह सुनिश्चित करती है कि डेटा से निकलने वाले निष्कर्ष सटीक और उपयोगी हों।

डेटा साइंस के 10 चरण (क्रम में):

1. **व्यावसायिक समझ (Business Understanding):** यह सबसे पहला और महत्वपूर्ण चरण है। इसमें व्यावसायिक दृष्टिकोण से समस्या, लक्ष्य और उद्देश्यों को परिभाषित किया जाता है। (उदाहरण के लिए: "ग्राहकों की संख्या क्यों कम हो रही है?")
2. **विश्लेषणात्मक दृष्टिकोण (Analytic Approach):** एक बार समस्या परिभाषित हो जाने के बाद, डेटा वैज्ञानिक सबसे अच्छे विश्लेषणात्मक रास्ते की पहचान करता है (जैसे: क्या हमें वर्गीकरण मॉडल (Classification) का उपयोग करना चाहिए या प्रतिगमन मॉडल (Regression) का?)।
3. **डेटा आवश्यकताएँ (Data Requirements):** परिभाषित समस्या को हल करने के लिए आवश्यक विशिष्ट डेटा (तारीखें, प्रारूप, वेरिएबल्स) की पहचान करना।
4. **डेटा संग्रह (Data Collection):** विभिन्न स्रोतों जैसे डेटाबेस, वेब स्क्रेपिंग या सर्वेक्षणों से डेटा एकत्र करना।

5. **डेटा की समझ (Data Understanding):** यह जाँचने के लिए कि क्या एकत्र किया गया डेटा वास्तव में समस्या का प्रतिनिधित्व करता है, वर्णनात्मक आंकड़ों (Descriptive statistics) और विज़ुअलाइज़ेशन का उपयोग करना।
6. **डेटा की तैयारी (Data Preparation):** यह **सबसे अधिक समय लेने वाला चरण** है। इसमें डेटा की सफाई करना, छूटे हुए मानों (Missing values) को संभालना और इसे मॉडलिंग के लिए उपयुक्त प्रारूप में बदलना शामिल है।
7. **एआई मॉडलिंग (AI Modelling):** पैटर्न खोजने या भविष्यवाणियां करने के लिए तैयार डेटा पर एल्गोरिदम (जैसे लीनियर रिग्रेशन या न्यूरल नेटवर्क) लागू करना।
8. **मूल्यांकन (Evaluation):** मॉडल का परीक्षण करना ताकि यह देखा जा सके कि क्या यह सटीक है और क्या यह मूल व्यावसायिक लक्ष्य को पूरा करता है।
9. **परिनियोजन (Deployment):** मॉडल को वास्तविक दुनिया के वातावरण (जैसे किसी ऐप या वेबसाइट) में उपयोग के लिए लागू करना।
10. **फीडबैक (Feedback):** आगे सुधार करने के लिए मॉडल वास्तविक दुनिया में कैसा प्रदर्शन कर रहा है, इस पर डेटा एकत्र करना।

Think of this unit as a **professional roadmap** for turning a big idea into a working AI project. Instead of just "guessing" how to build something, this chapter teaches you the exact system professionals use at companies like IBM.

Here are the objectives explained in a simple, everyday way:

1. The "Big Picture" Roadmap

The main goal is to show you that Data Science isn't just about writing code; it's a **10-step process**. You'll learn how to start with a real-world question (like "How do we reduce traffic?") and follow a specific path until you have a solution that actually works.

2. Asking the Right Questions

You will learn that before you touch any data, you have to clearly define the problem. It's the difference between saying "I want to build an app" and "I want to build an app that helps local farmers find the best market prices."

3. The "Garbage In, Garbage Out" Rule

A major objective is understanding how to pick, clean, and organize your data. If you feed a machine bad information, you'll get a bad result. You'll learn how to prepare your "ingredients" before you start "cooking" the model.

4. Checking Your Work (Validation & Metrics)

Building a model is only half the battle. This chapter teaches you how to prove that your AI is actually smart.

- **Validation:** Making sure the model works on new data, not just the data it already saw.
- **Metrics:** Using "scorecards" (like Accuracy or Precision) to put a number on how well the model is performing.

5. Completing the "Capstone" Project

Finally, the objective is to give you the tools to finish your **Capstone Project**. By the end, you won't just have a random script; you'll have a professional project built on a methodology that experts actually use.

In short: This chapter turns you from someone who "plays with data" into a **Data Scientist** who follows a reliable, professional system to solve problems.

Question and Answers

Q. 1.

Statement – I: A capstone project is the final project of an academic program.
Statement – II: For a capstone project, the students research on a topic independently.

(A) **Both Statement I and Statement II are correct.**
(B) Both Statement I and Statement II are incorrect.
(C) Statement I is correct but Statement II is incorrect.
(D) Statement II is correct but Statement I is incorrect.

Q. 2 Which of the following is NOT one of the three main stages of an AI project's life-cycle?

(A) Project scoping
(B) **Data visualization**
(C) Design or build phase
(D) Deployment in production

Q.3 At high level, any AI project follows six steps. What is the first step in the AI project process?

(A) Data gathering
(B) AI model construction
(C) **Problem definition**
(D) Deployment

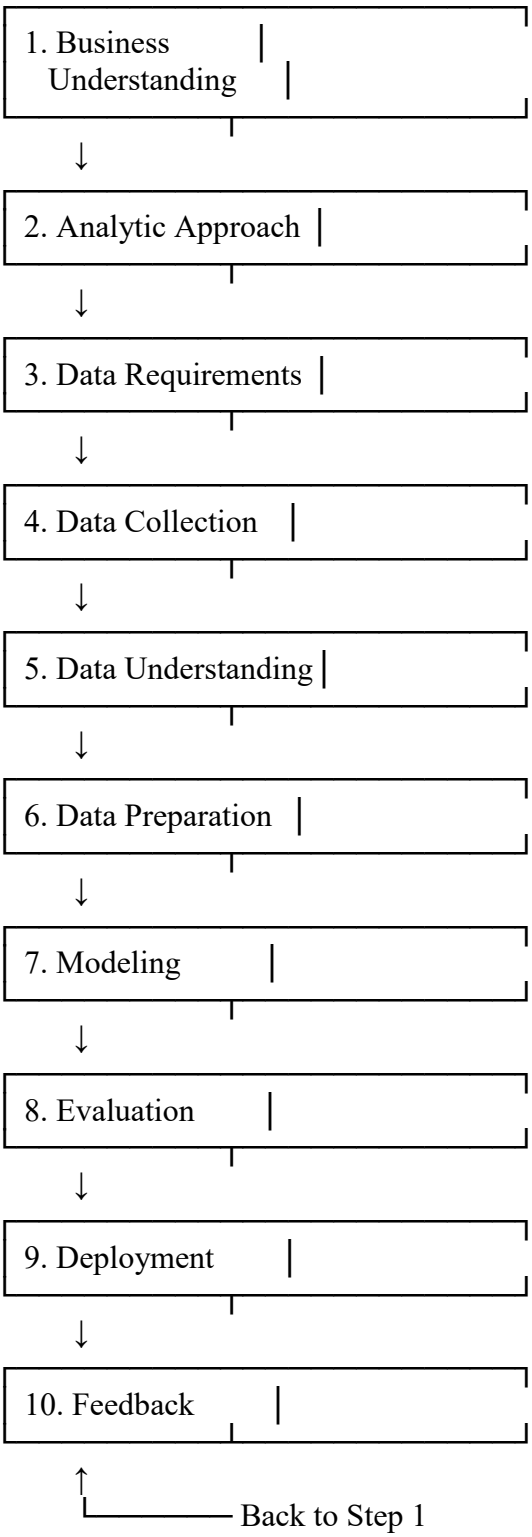
Q. 4 Which of the following is true about Design or Build stage of an AI Project Life-cycle?

(A) It is the first stage.
(B) **It is an iterative process.**
(C) It takes 15 days to complete.
(D) It includes only data acquisition and exploration.

Q. 5 "The _____ stage of an AI project cycle involves the planning and motivational aspects of your project.
(A) **Design**
(B) Deployment
(C) Testing
(D) Requirement Analysis"

Q. 6 "Which of the following analytical approach should be used to show relationships?
(A) Predictive
(B) Classification
(C) **Descriptive**
(D) Statistical Analysis"

Q. 7 "Draw a diagram to represent foundational methodology for Data Science."?
Ans:





Analytic Approach

This stage involves asking clarifying questions to stakeholders so that the data scientist can choose the right method. This stage does not include data collection. It focuses on selecting the methods of analysis and identifying the type of data required.

The team asks specific questions about the data purpose

The different questions that can be asked now are:

1. Do I need to find how much or how many? (Regression)

2. Which category does the data belong to? (Classification)

3. Can the data be grouped? (Clustering)

4. Is there any unusual pattern in the data? (Anomaly detection)

5. Which option should be given to the customer? (Recommendation)

To solve a particular problem , four main types of data analytics approaches are used. The represent different approaches to analyzing data. The analytic approach stage does not include data collection. It focuses on selecting the method of analysis and identifying the types of data required.

Types of data analytics

Descriptive Analytics: This summarizes past data to understand what has happened. It is the first step undertaken in data analytics to describe the trends and patterns using tools like graphs, charts etc. and statistical measures like mean, median, mode to understand the central tendency. This method also examines the spread of data using range, variance and standard deviation.

For example: To calculate the average marks of students in an exam or analyzing sales data from the previous year.

Diagnostic Analytics: It helps to understand the reason behind why some things have happened. This is normally done by analyzing past data using techniques like root cause analysis, hypothesis testing, correlation analysis etc. The main purpose is to identify the causes or factors that led to a certain outcome.

For example: If the sales of a company dropped, diagnostic analysis will help to find the cause for it, by analyzing questions like “Is it due to poor customer service” or “low product quality” etc.

Predictive Analytics: This uses the past data to make predictions about future events or trends, using techniques like regression, classification, clustering etc. Its main purpose is to foresee future outcomes and make informed decisions. For example: A company can use predictive analytics to forecast its sales, demand, inventory, customer purchase pattern etc., based on previous sales data.

Prescriptive Analytics: This recommends the action to be taken to achieve the desired outcome, using techniques such as optimization, simulation, decision analysis etc. Its purpose is to guide decisions by suggesting the best course of action based on data analysis. For example: To design the right strategy to increase the sales during festival season by analyzing past data and thus optimize pricing, marketing, production etc.

Types of Analytics VERY IMPORTANT

 1. Descriptive Analytics

English:

- Answers: What happened?
- Uses past data
- Example: Average marks

हिंदी:

क्या हुआ?

पिछले डेटा का विश्लेषण

2. Diagnostic Analytics

English:

- Answers: Why did it happen?
- Finds causes
- Example: Why sales decreased

हिंदी

क्यों हुआ?

कारण पता करना

3. Predictive Analytics

English:

- Answers: What will happen?
- Uses ML models
- Example: Predict exam score

हिंदी

क्या होगा?

भविष्य की भविष्यवाणी

4. Prescriptive Analytics

English:

- Answers: What should we do?
- Suggests actions

हिंदी:

क्या करना चाहिए?

निर्णय लेने में मदद

- Example: Best strategy for profit

2. From Requirements to Collection

Data requirements

What are the data requirements?

The requirements of data are determined by the analytic approach chosen in the previous stage. The 5W1H questioning method can be employed in this stage also to determine the data requirements. It is necessary to identify the data content, formats, and sources for initial data collection, in this stage.

5 h1w

Who , What, Where, When, Why

How

Determining the specific information needed for our analysis or project includes:

● identifying the types of data required, such as numbers, words, or images.

● considering the structure in which the data should be organized, whether it is in a table, text file, or database.

● identifying the sources from which we can collect the data, and

● any necessary cleaning or organization steps required before beginning the analysis.

This stage involves defining our data requirements, including the type, format, source, and necessary preprocessing steps to ensure the data is usable and accurate for our needs.

Data for a project can be categorized into three types: structured data (organized in tables, e.g., customer databases),

unstructured data (without a predefined structure, e.g., social media posts, images), and

semi-structured data (having some organization, e.g., emails, XML files).

Understanding these data types is essential for effective data collection and management in project development

Data collection

What occurs during data collection?

Data collection is a systematic process of gathering observations or measurements. In this phase, the data requirements are revised and decisions are made as to whether the collection requires more or less data.

→ Primary data Source

A primary data source refers to the original source of data, where the data is collected firsthand through direct observation, experimentation, surveys, interviews, or other methods. This data is raw, unprocessed, and unbiased, providing the most accurate and reliable information for research, analysis, or decision-making purposes. Examples include marketing campaigns, feedback forms, IoT sensor data etc.

→ Secondary data Source

A secondary data source refers to the data which is already stored and ready for use. Data given in books, journals, websites, internal transactional databases, etc. can be reused for data analysis. Some methods of collecting secondary data are social media data tracking, web scraping, and satellite data tracking. Some sources of online data are data.gov, World bank open data, UNICEF, open data network, Kaggle, World Health Organization, Google etc. Smart forms are an easy way to procure data online.

The Data Collection stage may be revisited after the Data Understanding stage, where gaps in the data are identified, and strategies are developed to either collect additional data or make substitutions to ensure data completeness.

□ Data Collection – Short Notes

- **Definition**: Systematic process of gathering observations or measurements to meet research or analysis requirements.
- **Process**:
  - Revise data requirements.
  - Decide if more or less data is needed.
  - May revisit after *Data Understanding* stage to fill gaps or substitute missing data.

□ Primary Data Source

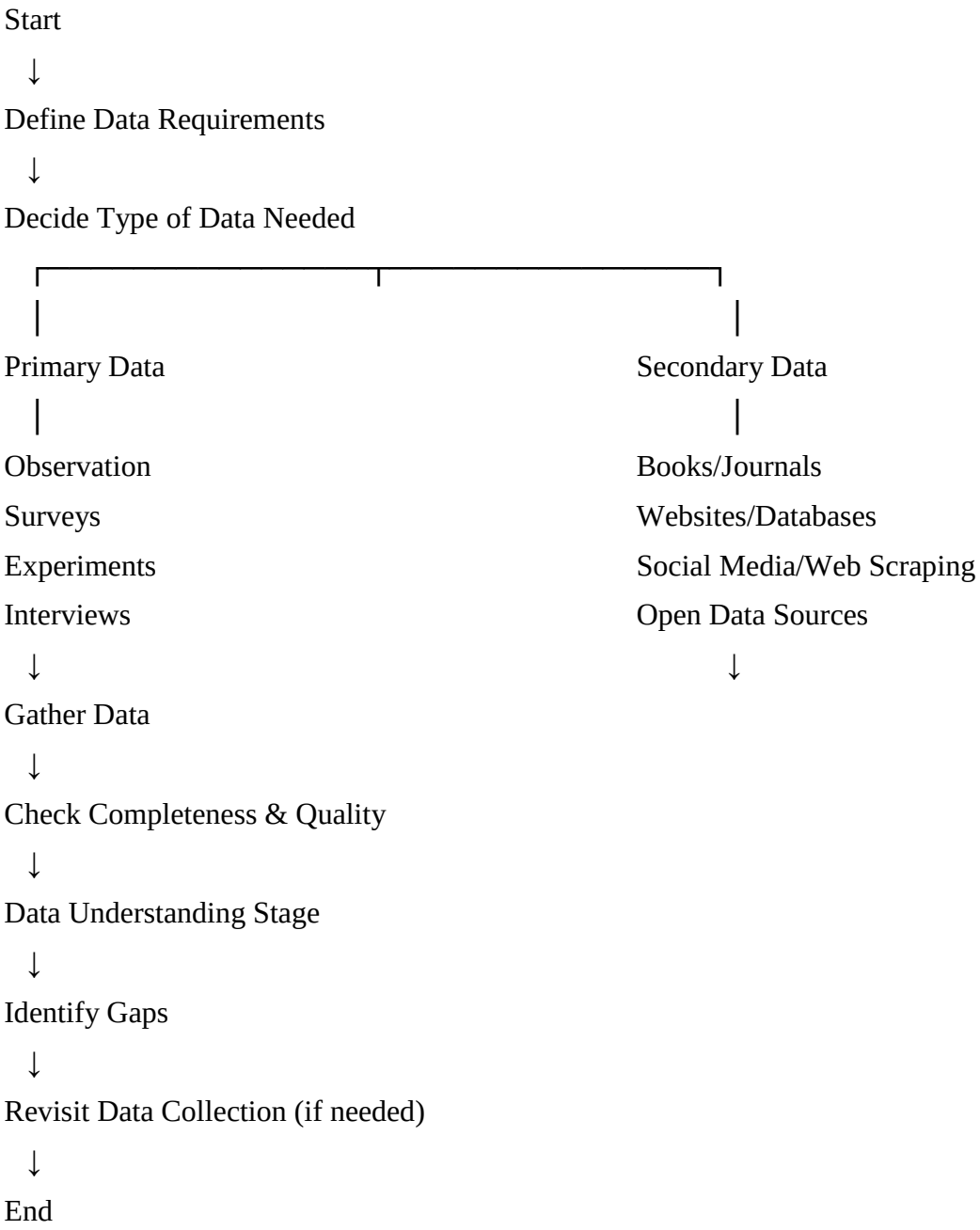
- **Meaning**: Original, firsthand data collected directly.
- **Methods**: Observation, experiments, surveys, interviews.
- **Nature**: Raw, unprocessed, unbiased.
- **Examples**: Marketing campaign results, feedback forms, IoT sensor readings.

□ Secondary Data Source

- **Meaning**: Pre-existing data already stored and available.
- **Sources**: Books, journals, websites, internal databases.
- **Methods**: Social media tracking, web scraping, satellite data.
- **Examples**: Data.gov, World Bank Open Data, UNICEF, WHO, Kaggle, Google datasets.

□ Key Point

- Data collection is iterative: after analyzing data, gaps may require additional collection or substitution to ensure completeness.


Start
Define Data Requirements
Decide Type of Data Needed
Primary Data
Secondary Data
Observation
Surveys
Experiments
Interviews
Books/Journals
Websites/Databases
Social Media/Web Scraping
Open Data Sources
Gather Data
Check Completeness & Quality
Data Understanding Stage
Identify Gaps
Revisit Data Collection (if needed)
End

3. From Understanding to Preparation

Data Understanding

Is the data collected representative of the problem to be solved?

Data Understanding encompasses all activities related to constructing the dataset. In this stage, we check whether the data collected represents the problem to be solved or not. The relevance, comprehensiveness, and suitability of the data for addressing the specific problem or question at hand are evaluated. Techniques such as descriptive statistics (univariate analysis, pairwise correlation etc.) and visualization (Histogram) can be applied to the dataset, to assess the content, quality, and initial insights about the data.

Data preparation

What additional work is required to manipulate and work with the data?

This stage covers all the activities to build the set of data that will be used in the modelling step. Data is transformed into a state where it is easier to work with.

Data preparation includes

● cleaning of data (dealing with invalid or missing values, removal of duplicate values and assigning a suitable format)

● combine data from multiple sources (archives, tables and platforms)

● transform data into meaningful input variables

Feature Engineering is a part of Data Preparation. The preparation of data is the most timeconsuming step among the Data Science stages.

Feature engineering is the process of selecting, modifying, or creating new features (variables) from raw data to improve the performance of machine learning models.

For example:

Suppose you're building an ML model to predict price of houses, and you have the following data:

● Raw Data: Area of the house (in sq.ft), number of bedrooms, and the year the house was built.

To improve the model's performance, you might create new features such as age of the house and price per square foot which can be derived from the raw data.

● New Features

1. Age of the house = Current year - Year built.

2. Price per square foot = Price of the house / Area.

These new features can help the model make more accurate predictions.

3. From Understanding to Preparation

□ Data Understanding

- **Definition**: Activities to evaluate whether collected data represents the problem to be solved.
- **Focus**: Relevance, comprehensiveness, and suitability of data.
- **Techniques**:
  - Descriptive statistics (univariate analysis, pairwise correlation).
  - Visualization (e.g., histograms).
- **Goal**: Assess content, quality, and gain initial insights.

□ Data Preparation

- **Definition**: Activities to transform raw data into a usable form for modeling.
- **Tasks**:
  - Cleaning (handle missing/invalid values, remove duplicates, format data).
  - Combining data from multiple sources (archives, tables, platforms).
  - Transforming data into meaningful input variables.
- **Note**: Most time-consuming stage in the data science process.

□ Feature Engineering (part of Data Preparation)

- **Definition**: Selecting, modifying, or creating new features to improve model performance.
- **Purpose**: Enhance predictive power of machine learning models.
- **Example** (House price prediction):
  - Raw features: Area, bedrooms, year built.
  - New features:
    1. Age of house = Current year − Year built.
    2. Price per square foot = Price ÷ Area.
- **Impact**: Helps models make more accurate predictions.

□ Key Insights

- **Data Understanding** ensures relevance and quality.
- **Data Preparation** makes raw data usable.
- **Feature Engineering** boosts model accuracy by adding meaningful variables.
- Together, these stages form the backbone before modeling and deployment.

4. From Modelling to Evaluation

AI modelling

In what way can the data be visualized to get to the required answer?

The modelling stage uses the initial version of the dataset prepared and focuses on developing models according to the analytical approach previously defined.

The modelling process is usually iterative, leading to the adjustments in the preparation of data. For a determined technique, Data scientists can test multiple algorithms to identify the most suitable model for the Capstone Project.

Data Modelling focuses on developing models that are either descriptive or predictive.

1. Descriptive Modeling: It is a concept in data science and statistics that focuses on summarizing and understanding the characteristics of a dataset without making predictions or decisions.

The goal of descriptive modeling is to describe the data rather than predict or make decisions based on it. This includes summarizing the main

characteristics, patterns, and trends that are present in the data.

Descriptive modeling is useful when you want to understand what is happening within your data and how it behaves, but not necessarily why it happens.

Common Descriptive Techniques:

- Summary Statistics: This includes measures like:
  - Mean (average), Median, Mode
  - Standard deviation, Variance
  - Range (difference between the highest and lowest values)
  - Percentiles (e.g., quartiles)
- Visualizations: Graphs and charts to represent the data, such as:
  - Bar charts
  - Histograms
  - Pie charts
  - Box plots
  - Scatter plots

2. Predictive modeling: It involves using data and statistical algorithms to identify patterns and trends in order to predict future outcomes or values. It relies on historical data and uses it to create a model that can predict future behavior or trends or forecast what might happen next. It involves techniques like regression, classification, and time-series forecasting, and can be applied in a variety of fields, from predicting exam scores to

forecasting weather or stock prices. While it is a powerful tool, students must also understand its limitations and the importance of good data.

The data scientist will use a training set for predictive modeling. A training set is a set of historical data in which the outcomes are already known. The training set acts like a gauge to determine if the model needs

to be calibrated. In this stage, the data scientist will play around with different algorithms to ensure that the variables selected are actually required.

4. From Modelling to Evaluation

AI Modelling – Short Notes

What is AI Modelling?

AI Modelling is the process of using prepared data to create models that help in understanding data or predicting future outcomes.

The modelling stage:

- Uses the prepared dataset
- Applies different algorithms
- Tests and improves models
- Helps answer the main problem/question

The process is usually **iterative**, meaning data scientists repeatedly improve the model and data preparation.

Types of Data Modelling

1. Descriptive Modelling

Definition

Descriptive modelling focuses on **understanding and summarizing data** without making predictions.

Purpose

- Finds patterns and trends
- Explains “what is happening” in the data
- Helps in data visualization and analysis

Common Descriptive Techniques

A. Summary Statistics

- Mean (Average)
- Median
- Mode
- Standard Deviation
- Variance
- Range
- Percentiles/Quartiles

B. Visualizations

- Bar Chart
- Histogram
- Pie Chart
- Box Plot
- Scatter Plot

Example

Analyzing students' marks to find:

- average score
- highest marks
- performance distribution

2. Predictive Modelling

Definition

Predictive modelling uses historical data and algorithms to **predict future outcomes**.

Purpose

- Forecast future trends
- Predict behavior or results
- Support decision-making

Techniques Used

- Regression
- Classification
- Time-Series Forecasting

Training Set

A **training set** is historical data with known outcomes used to train and test the model.

Example

- Predicting exam scores
- Weather forecasting
- Stock price prediction

Important Point

Good quality data is necessary for accurate predictions.

Difference Between Descriptive and Predictive Modelling

Descriptive Modelling	Predictive Modelling
Descriptive Modelling	Predictive Modelling
Explains current data	Predicts future outcomes
Focuses on patterns/trends	Focuses on forecasting
No prediction involved	Uses prediction algorithms
Example: Average marks	Example: Predicting final result

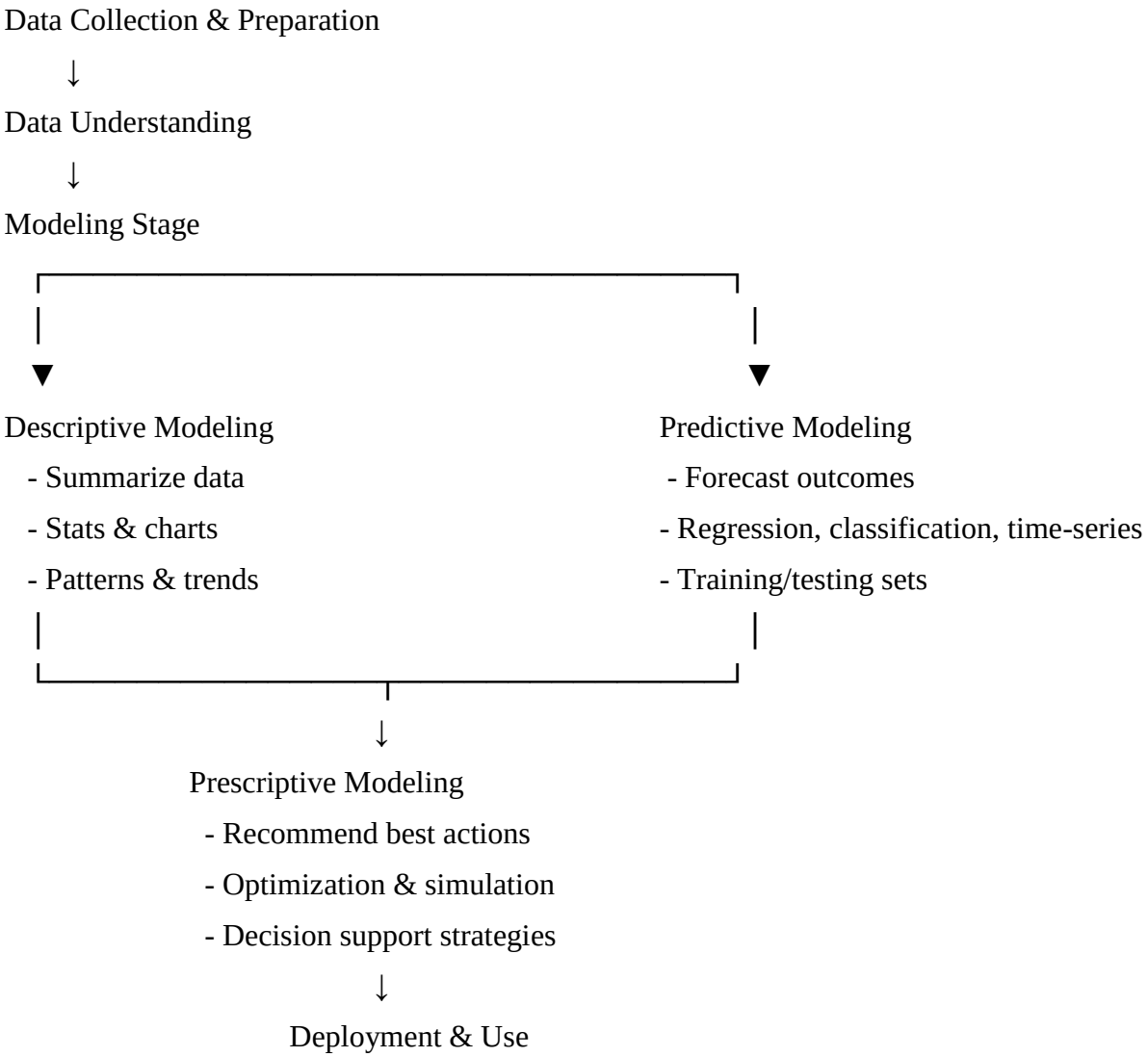
 Descriptive vs Predictive vs Prescriptive Modeling

Aspect	Descriptive Modeling	Predictive Modeling	Prescriptive Modeling
Purpose	Summarize and understand dataset characteristics	Forecast future outcomes based on historical data	Recommend actions to achieve desired outcomes
Focus	*What is happening*	*What will happen*	*What should we do*
Techniques	Summary statistics, visualization (bar charts, histograms, box plots)	Regression, classification, time-series forecasting	Optimization, simulation, decision analysis
Data Used	Current dataset	Historical dataset (training + test sets)	Predictive model outputs + business constraints
Outcome	Insights, patterns, trends	Predictions, forecasts	Actionable strategies, best decisions
Applications	Customer behavior analysis, survey summaries	Exam score prediction, weather forecasting, stock price prediction	Supply chain optimization, personalized recommendations, resource allocation

Key Insight

- **Descriptive** → Understand the present.
- **Predictive** → Anticipate the future.
- **Prescriptive** → Decide the best course of action.

Together, they form a progression: **Describe → Predict → Prescribe**, moving from insight to foresight to decision-making.



Evaluation

Does the model used really answer the initial question or does it need to be adjusted?

Evaluation in an AI project cycle is the process of assessing how well a model performs after training. It involves using test data to measure metrics like accuracy, precision, recall, or F1 score. This helps determine if the model is reliable and effective before deploying it in real-world situations.

Model evaluation can have two main phases.

First phase – Diagnostic measures

It is used to ensure the model is working as intended.

If the model is a **predictive model**, a decision tree can be used to evaluate the output of the model, check whether it is aligned to the initial design orrequires any adjustments.

If the model is a **descriptive model**, one in which relationships are being assessed, then a testing set with known outcomes can be applied, and the model can be refined as needed.

Second phase – Statistical significance test

This type of evaluation can be applied to the model to verify that it accurately processes and interprets the data. This is designed to avoid unnecessary second guessing when the answer is revealed

? Model Evaluation

- **Purpose**: To assess how well a trained model performs before deployment.
- **Key Metrics**: Accuracy, precision, recall, F1 score.
- **Goal**: Ensure reliability, effectiveness, and alignment with the initial design.

Phase 1 – Diagnostic Measures

- **Predictive Models**: Use tools like decision trees to check if outputs match design expectations.
- **Descriptive Models**: Apply test sets with known outcomes to verify relationships and refine the model.
- **Focus**: Detect errors early and adjust the model if needed.

Phase 2 – Statistical Significance Test

- **Purpose**: Verify that the model's results are not due to chance.
- **Method**: Apply statistical tests to confirm the model interprets data correctly.
- **Outcome**: Builds confidence in the model, reduces unnecessary second-guessing.

 Summary: Evaluation ensures the model answers the initial question effectively. Diagnostic checks confirm functionality, while statistical tests validate reliability. Together, they determine if the model is ready for real-world use or needs adjustment.

the **difference between Model Validation and Model Evaluation**:

Model Validation

- **Definition**: The process of checking whether the model is built correctly and meets the design requirements during training.
- **Focus**: Ensures the model generalizes well to unseen data and avoids overfitting.
- **Methods**:
  - Splitting data into training, validation, and test sets.
  - Cross-validation techniques (e.g., k-fold).
- **Goal**: Confirm the model is *fit for purpose* before final testing.

Model Evaluation

- **Definition**: The process of assessing how well the trained model performs after training, using test data.
- **Focus**: Measures performance with metrics like accuracy, precision, recall, F1 score, ROC curve, etc.
- **Phases**:
  - Diagnostic checks (does the model answer the initial question?).
  - Statistical significance tests (are results reliable and not due to chance?).
- **Goal**: Determine if the model is *ready for deployment* in real-world scenarios.

⚖ Key Difference

Aspect	Model Validation	Model Evaluation
Stage	During training	After training
Purpose	Ensure correctness & generalization	Assess performance & reliability
Data Used	Validation set	Test set
Outcome	Model refinement	Deployment decision

 Summary:

- **Validation** = "Am I building the model right?"
- **Evaluation** = "Does the model actually work well in practice?"

5. From Deployment to Feedback

Deployment

How does the solution reach the hands of the user?

Deployment refers to the stage where the trained AI model is made available to the users in real-world applications.

Data scientists must make the stakeholders familiar with the tool produced in different scenarios.

Once the model is evaluated and the data scientist is confident it will work, it is deployed and put to the ultimate test.

Depending on the purpose of the model, it may be rolled out to a limited group of users or in a test environment, to build up confidence in applying the outcome for use across the board.

Deploying a model into a live business process, frequently necessitates the involvement of additional internal teams, skills and technology.

Feedback

Is the problem solved? Has the question been satisfactorily answered?

The last stage in the methodology is feedback.

This includes results collected from the deployment of the model, feedback on the model's performance from the users and clients, and observations from how the model works in the deployed environment.

This process continues till the model provides satisfactory and acceptable results.

Feedback from the users will help to refine the model and assess it for performance and impact.

The process from modelling to feedback is highly iterative.

Data Scientists may automate any or all of the feedback so that the model refresh process speeds up and can get quick improved results.

Feedback from users can be received in many ways.



Deployment

- **Purpose**: To make the trained AI model available for real-world use.
- **Approach**: Roll out to limited users or test environments first, then scale up.
- **Goal**: Ensure the model integrates smoothly into business processes and delivers expected outcomes.

Phase 1 – Rollout Strategy

- **Limited Deployment**: Test with small user groups to build confidence.
- **Full Deployment**: Expand once reliability is proven.
- **Focus**: Collaboration with internal teams, ensuring technical readiness.

Phase 2 – Integration into Processes

- **Requirement**: Additional skills, technology, and support teams.
- **Method**: Embed the model into live workflows.
- **Outcome**: Real-world validation of the model's effectiveness.

Summary: Deployment is the ultimate test of the model. Careful rollout and integration ensure it works as intended before scaling widely.

Feedback

- **Purpose**: To evaluate if the model solves the problem and meets user needs.
- **Sources**: User responses, client observations, system performance data.
- **Goal**: Refine the model iteratively for better accuracy and impact.

Phase 1 – User & Client Feedback

- **Channels**: Support tickets, surveys, social likes, website feedback, call centers, in-store feedback, local media.
- **Focus**: Gather diverse insights on performance and usability.

Phase 2 – Iterative Improvement

- **Process**: Automate feedback collection where possible.
- **Method**: Use feedback loops to retrain and refresh the model quickly.
- **Outcome**: Continuous improvement until results are satisfactory.

Summary: Feedback ensures the model evolves. It validates performance, highlights gaps, and drives refinements. The cycle of deployment → feedback → improvement makes AI solutions sustainable.

MODEL VALIDATION

Evaluating the performance of a trained machine learning model is essential. Model Validation offers a systematic approach to measure its accuracy and reliability, providing insights into how well it generalizes to new, unseen data.

Model validation is the step conducted post Model Training, wherein the effectiveness of the trained model is assessed using a testing dataset. Validating the machine learning model during the training and development stages is crucial for ensuring accurate predictions. The benefits of Model Validation include

• Enhancing the model quality.

• Reduced risk of errors

• Prevents the model from overfitting and underfitting.

Model Validation Techniques

The commonly used Validation techniques are Train-test split, K-Fold Cross Validation, Leave One out Cross Validation, Time Series Cross Validation etc. Let's discuss the Train test split and K-Fold Cross Validation

Train Test Split

The train-test split is a technique for evaluating the performance of a machine learning algorithm. It can be used for classification or regression problems and can be used for any supervised learning algorithm

The procedure involves taking a dataset and dividing it into two subsets, as shown in Fig 2.4. The first subset is used to fit/train the model and is referred to as the training dataset. The second subset is used to test the model. It is with the testing data that predictions are made and compared to the expected values. This second dataset is referred to as the test dataset.

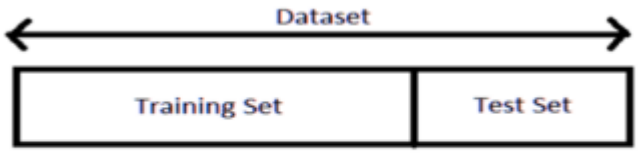


Fig 2.4

Train Dataset: Used to fit the machine learning model.

Test Dataset: Used to evaluate the fit machine learning model.

The objective is to estimate the performance of the machine learning model on new data, i.e., data not used to train the model.

This is how we expect to use the model in practice. The procedure is to fit it on available data with known inputs and outputs, then make predictions on new examples in the future where we do not have the expected output or target values.

The train-test procedure is appropriate when there is a sufficiently large dataset available.

How to Configure the Train-Test Split

The procedure has one main configuration parameter, which is the size of the train and test sets. This is most commonly expressed as a percentage between 0 and 1 for either the train or test datasets. For example, a training set with the size of 0.67 (67 percent) means that the remaining percentage 0.33 (33 percent) is assigned to the test set. There is no optimal split percentage.

Nevertheless, common split percentages include:

- Train: 80%, Test: 20%
- Train: 70%, Test: 30%
- Train: 67%, Test: 33%

K-Fold Cross Validation

Cross Validation is a technique used to evaluate a model's performance. It splits the data into multiple parts or folds. It trains the model on some folds and tests it on other folds and repeats this process for a number fixed by the data scientist. In cross-validation, we run our modeling process on different subsets of the data to get multiple measures of model quality.

k-fold cross validation

In k-fold cross validation we will be working with k subsets of datasets. For example, if we divide the data into 5 folds or 5 pieces, as shown in Fig 2.5, each being 20% of the full dataset, then k=5.

We run an experiment called experiment 1 which uses the first fold as a holdout set (validation), and remaining four folds as training data. This gives us a measure of model quality based on a 20% holdout set. We then run a second experiment, where we hold out data from the second fold. This gives us a second estimate of model quality. We repeat this process, using every fold once as the holdout. Putting this together, 100% of the data is used as a holdout at some point.

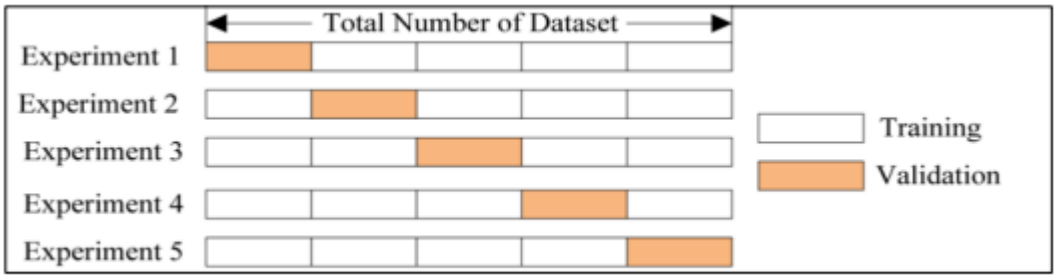


Fig 2.5

Cross-validation gives a more accurate measure of model quality, which is especially important if you are making a lot of modelling decisions. However, it can take more time to run, because it estimates models once for each fold.

Difference between Train-Test Split and Cross Validation

Train-Test Split	Cross Validation
Normally applied on large datasets	Normally applied on small datasets
Divides the data into training data set and testing dataset.	Divides a dataset into subsets (folds), trains the model on some folds, and evaluates its performance on the remaining data.
Clear demarcation on training data and testing data.	Every data point at some stage could be in either testing or training data set.