WEEK 1: AI FUNDAMENTALS AND PYTHON BASICS

DAY 3 (25/06/2025)

Understanding the AI Project Cycle:

Today's session was focused on revisiting and deeply understanding the **AI Project Cycle** — the step-by-step process that every Data Scientist follows when designing, building, and testing an AI system. The goal was to move beyond simply knowing the stages and to learn how and why each stage contributes to a successful AI project.

1. Problem Scoping

This is the **initial and most crucial non-technical phase**, where we clearly define what problem we are solving and why it matters. Before jumping into data or coding, we must understand the context and objectives behind the AI solution.

To structure our understanding, we use the **4 Ws** framework:

- **WHO** are the people who will benefit from the solution (the target users or audience)?
- **WHAT** is the problem we are solving, and how do we measure that it actually is a problem?
- **WHERE** does this problem occur (specific location, conditions, or scenario)?
- **WHY** do we want to solve this problem (the benefit, purpose, or potential impact)?

2. Data Collection

Once the problem is clearly defined, the next step is to **collect relevant data** — the raw material that fuels any AI system. The quality and quantity of data directly affect how well the model can learn.

Depending on the project, data can take various forms:

- **Images or videos** (used in visual recognition system).
- **Text data** (used in chatbots or language models).
- **Numerical tables or sensor readings** (used in finance, healthcare, or IoT applications).

It is important to ensure that the collected data is representative, diverse, and free from bias. For instance, if a waste classification dataset contains only plastic and paper but no metal or glass, the model will fail in real-life conditions. Thus, careful data sourcing is key to reliable AI performance.

3. Data Preprocessing

Raw data is rarely clean or ready for analysis. It often contains **missing values**, **duplicates**, **noise**, or **inconsistent formats**.
Data preprocessing is the step where we **clean, transform, and prepare** the data to make it suitable for model training.

Common preprocessing tasks include:

- **Handling Missing Values:** Filling in gaps using statistical methods (mean, median, mode) or removing incomplete records.
- **Standardization and Normalization:** Scaling features so they are comparable and within a similar numerical range.
- **Encoding Categorical Variables:** Converting text-based labels (e.g., “plastic,” “metal,” “paper”) into numerical representations that the machine can understand.

This step might seem tedious, but it’s critical — clean data often leads to models that learn faster and perform more accurately.

4. Data Modelling

This is the heart of the AI project, where real learning takes place.
Here, we select the **appropriate algorithm or model** based on the type of data and the problem we are solving. For example:

- **Neural Networks** or **Convolutional Neural Networks (CNNs)** for image recognition tasks.
- **Regression models** for numerical prediction (like predicting house prices).
- **Decision Trees or Random Forests** for classification problems.

Once the model is selected, the **training process** begins. The AI learns from the input data by identifying hidden patterns, relationships, and dependencies. This is where the system builds its internal "understanding" of the data to make predictions or classifications in the future.

5. Model Evaluation

After training, we must ensure the model actually works well in real-world conditions. This is done through **evaluation**, where we test the model's performance on **unseen or new data** that was not part of the training set.

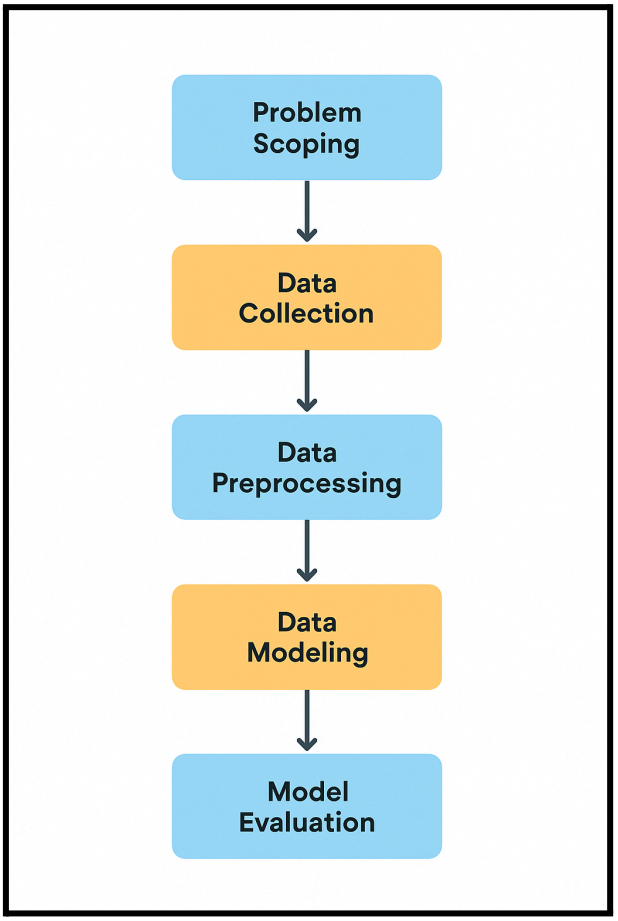
Evaluation gives us an unbiased estimate of how the model will perform on real data. Several metrics are used to measure this performance:

- **Accuracy:** The proportion of correct predictions.
- **Precision and Recall:** Precision measures correctness among positive predictions, while Recall measures how many actual positives were captured.
- **F1-Score:** A balance between Precision and Recall — helpful when dealing with imbalanced datasets.

A good model not only achieves high accuracy but also performs consistently across different data types. Evaluation is also where we identify errors, tune parameters, and improve performance before deploying the model.

Reflection

Today's session was very insightful because it connected every stage of the AI project lifecycle into a logical sequence. I understood that being a good Data Scientist is not only about knowing algorithms but also about **following a structured workflow** — from defining the right problem to validating the final model.



AI Project Lifecycle