WEEK 2: MACHINE LEARNING CONCEPTS

DAY 8 (02/07/2025)

Machine Learning Workflow and Data Preprocessing:

Machine Learning doesn't just involve building models—it's a complete workflow that starts with raw data and ends with meaningful predictions. The quality of each stage determines how effective the final model will be.

1. Machine Learning Workflow

A typical ML workflow follows these key steps:

1. **Data Collection:**
   Gathering data from sources like databases, sensors, etc.

2. **Data Preprocessing:**
   Cleaning and transforming raw data into a structured, usable format.

3. **Feature Engineering:**
   Selecting and creating input features that help the model learn better.

4. **Model Training:**
   Feeding the processed data into algorithms so that they can learn patterns.

5. **Model Validation:**
   Testing the model on unseen data to ensure it generalizes well.

6. **Model Evaluation:**
   Measuring how accurate and reliable the model is using metrics.

7. **Deployment and Monitoring:**
   Integrating the model into applications and monitoring its performance over time.

2. Data Preprocessing

Real-world data is messy — it may contain **missing values, outliers, or irrelevant information**. Preprocessing makes it clean and consistent.

Key steps include:

- **Handling Missing Values:** Replace missing entries with mean/median or remove them.
- **Encoding Categorical Data:** Convert text labels (like "Male" or "Female") into numerical codes for algorithms.
- **Feature Scaling:** Normalize or standardize values so that all features are on a similar scale.
- **Outlier Removal:** Identify and remove abnormal data points that can distort model training.

Example:

Suppose we're predicting Body Mass Index (BMI) using *"height in cm"* (values around 150–190) and *"age in years"* (10–70).

If not scaled, the algorithm might give more weight to height simply because it has higher numerical values, even though both features are important.

3. Data Splitting: Training, Validation & Testing Sets

Before training, the dataset is divided into three parts:

- **Training Set (70–80%)** → Used to train the model.
- **Validation Set (10–15%)** → Used to tune parameters and prevent overfitting.
- **Testing Set (10–15%)** → Used to evaluate final model performance.

4. Model Validation

Model validation helps us check how well a machine learning model performs on **new data**. It ensures that the model has **actually learned patterns** and is not just **memorizing** the training examples.

To validate a model, we usually split the dataset into two main parts:

- **Training Set:** Used to teach the model.
- **Testing Set:** Used to check how well the model performs on unseen data.

If the model performs well on the training data but poorly on the testing data, it means the model is **overfitting** — it has memorized the data instead of learning from it.
If it performs poorly on both, it is **underfitting** — it hasn't learned enough.

Example:

If we train a model to predict student marks using 80% of the data and test it on the remaining 20%, good performance on the test set means the model has learned to **generalize** properly and can predict marks for new students too.

Reflection

Today I learned that machine learning is more than building models—it's a full workflow from data collection to deployment. Preprocessing data, like handling missing values and scaling features, is crucial for accurate predictions. Splitting data and validating models helps prevent overfitting and ensures the model generalizes well to new data.