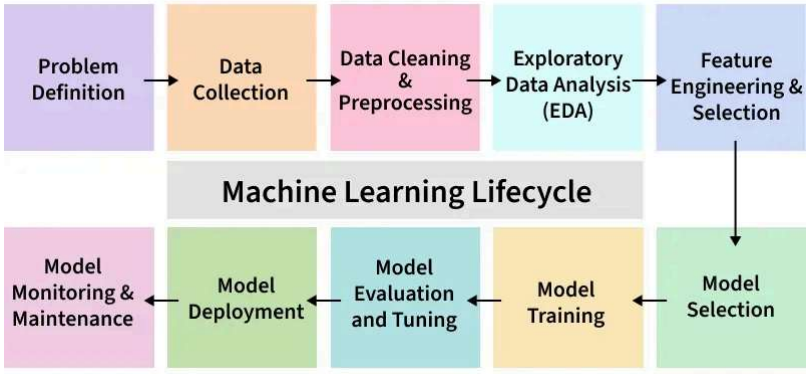
6.01: Introduction to EDA & Feature Engineering

Exploratory Data Analysis (EDA) and Feature Engineering are two of the most important steps in the machine learning workflow.
They help transform raw data into meaningful insights and build a strong foundation for accurate and robust models.

Machine Learning Lifecycle



Exploratory Data Analysis (EDA)

EDA is the process of exploring, understanding, and summarizing datasets before applying any machine learning algorithms.
It helps in:

- Identifying data patterns and trends
- Detecting outliers and missing values
- Understanding variable distributions
- Building intuition for model selection

Common techniques in EDA:

- Descriptive statistics (mean, median, mode, variance, etc.)
- Data visualization (histograms, scatter plots, box plots, heatmaps)
- Correlation and covariance analysis
- Handling missing values and outliers

Feature Engineering

Feature Engineering is the process of creating, transforming, and selecting features to improve model performance.

It bridges the gap between raw data and machine learning algorithms.

Key steps in Feature Engineering:

- Handling categorical variables (encoding techniques)
- Feature scaling (normalization & standardization)
- Creating new features from existing ones
- Feature selection and dimensionality reduction (PCA, Lasso, etc.)
- Domain knowledge-driven transformations

Why are EDA & Feature Engineering important?

- Improve model accuracy and reliability
- Reduce noise and redundancy in data
- Enhance interpretability of models
- Save time and resources during training

Mastering EDA and Feature Engineering is essential for every data scientist and ML practitioner, as it often determines the success of your machine learning project.

🎉 Happy Learning !

Team DecodeAiML