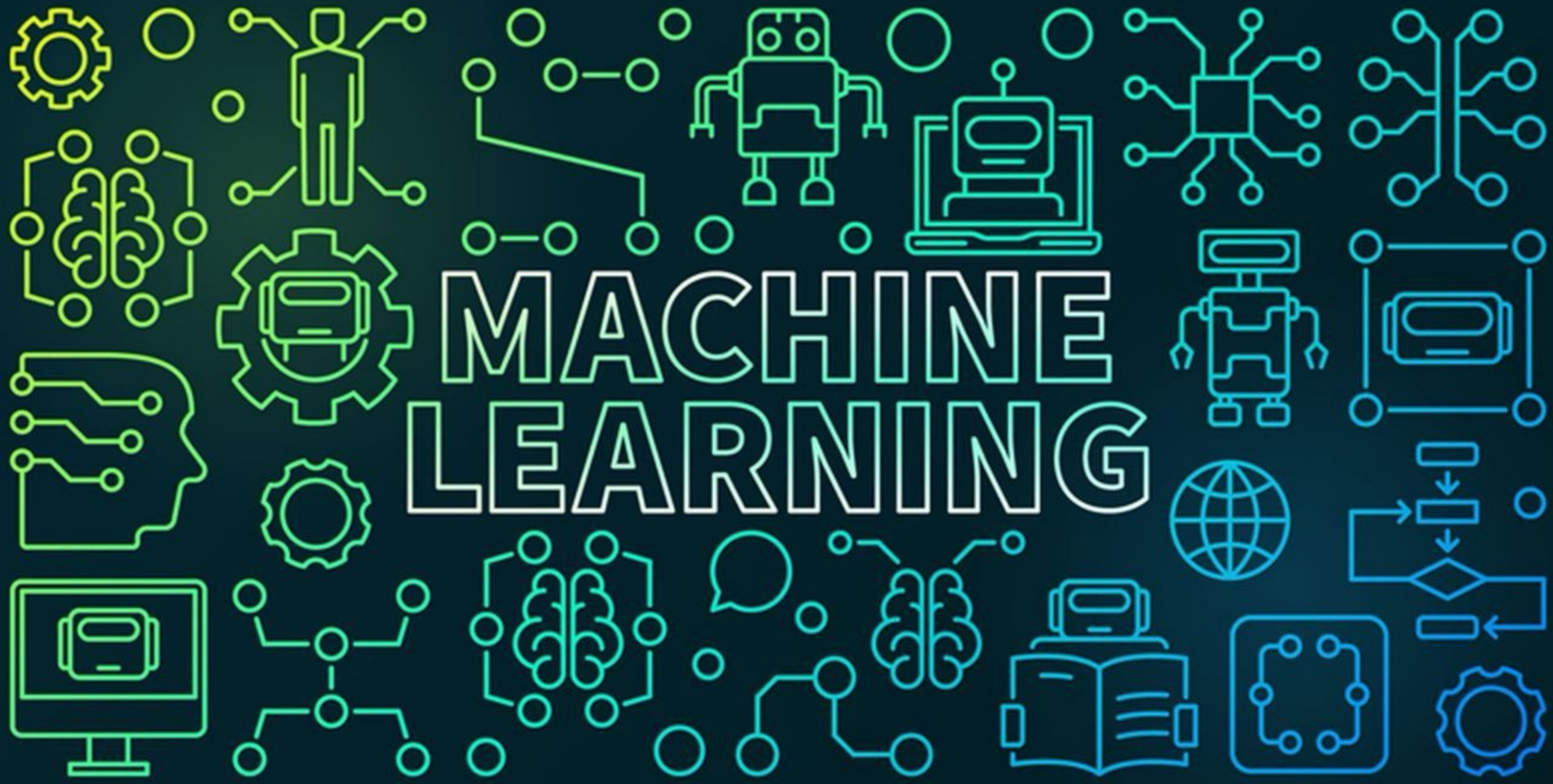
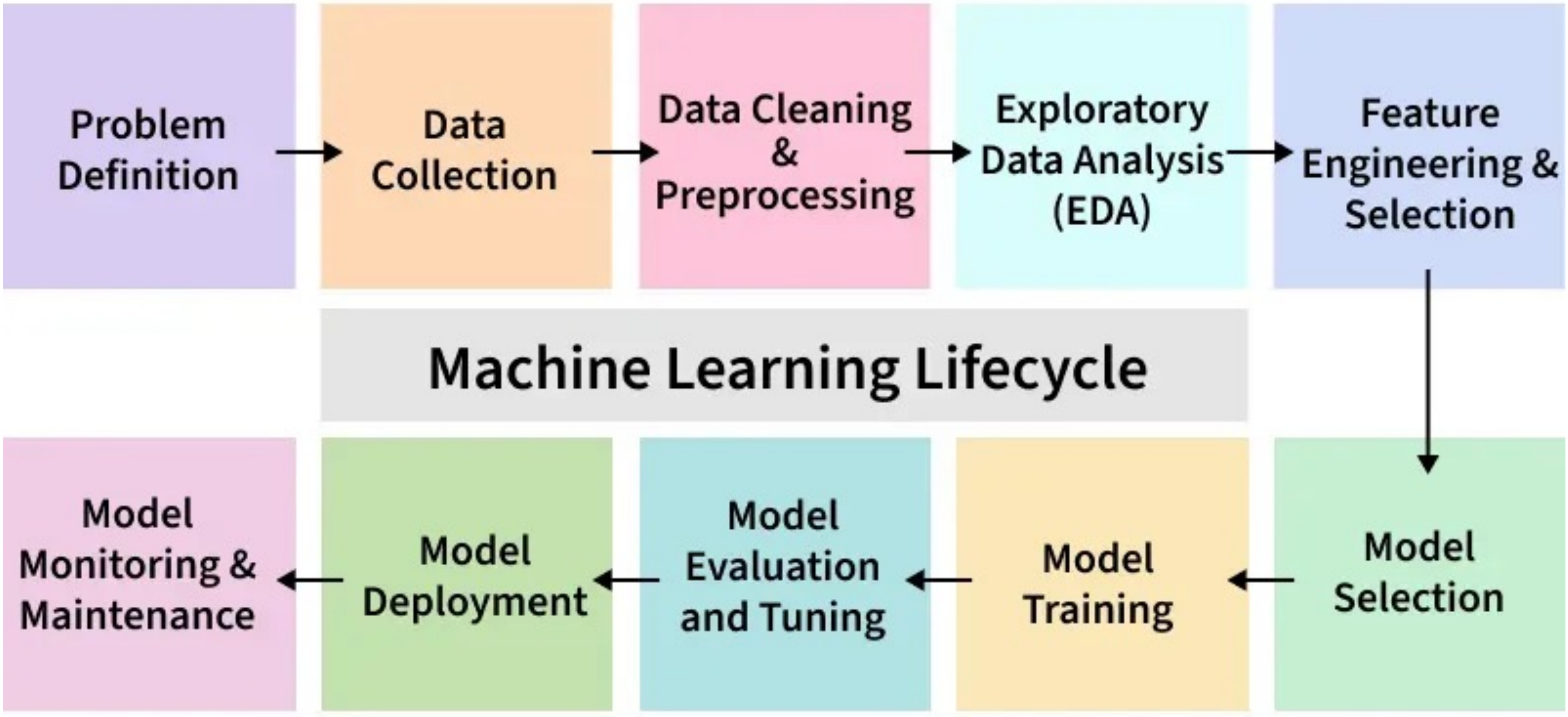


Muhammad Atif Saeed (Lecturer DS & AI)

Machine Learning Lifecycle



Machine Learning Lifecycle

Problem Definition → Data Collection → Data Cleaning & Preprocessing → Exploratory Data Analysis (EDA) → Feature Engineering & Selection → Model Selection → Model Training → Model Evaluation and Tuning → Model Deployment → Model Monitoring & Maintenance

Step 1: Problem Definition

- The first step is identifying and clearly defining the business problem. A well-framed problem provides the foundation for the entire lifecycle. Important things like project objectives, desired outcomes and the scope of the task are carefully designed during this stage.
  - Collaborate with stakeholders to understand business goals
  - Define project objectives, scope and success criteria
  - Ensure clarity in desired outcomes

Step 2: Data Collection

- Data Collection phase involves systematic collection of datasets that can be used as raw data to train model. The quality and variety of data directly affect the model's performance.
- Here are some basic features of Data Collection:
  - Relevance: Collect data should be relevant to the defined problem and include necessary features.
  - Quality: Ensure data quality by considering factors like accuracy and ethical use.
  - Quantity: Gather sufficient data volume to train a robust model.
  - Diversity: Include diverse datasets to capture a broad range of scenarios and patterns.

Step 3: Data Cleaning and Preprocessing

- Raw data is often messy and unstructured and if we use this data directly to train then it can lead to poor accuracy. We need to do data cleaning and preprocessing which often involves:
  - Data Cleaning: Address issues such as missing values, outliers and inconsistencies in the data.
  - Data Preprocessing: Standardize formats, scale values and encode categorical variables for consistency.
  - Data Quality: Ensure that the data is well-organized and prepared for meaningful analysis.

Step 4: Exploratory Data Analysis (EDA)

- Exploration: Use statistical and visual tools to explore patterns in data.
- Patterns and Trends: Identify underlying patterns, trends and potential challenges within the dataset.
- Insights: Gain valuable insights for informed decisions making in later stages.
- Decision Making: Use EDA for feature engineering and model selection.

Step 5: Feature Engineering and Selection

- Feature engineering and selection is a transformative process that involve selecting only relevant features to enhance model efficiency and prediction while reducing complexity.
- Here are the basic features of Feature Engineering and Selection:
  - Feature Engineering: Create new features or transform existing ones to capture better patterns and relationships.
  - Feature Selection: Identify subset of features that most significantly impact the model's performance.
  - Domain Expertise: Use domain knowledge to engineer features that contribute meaningfully for prediction.
  - Optimization: Balance set of features for accuracy while minimizing computational complexity.

Step 6: Model Selection

- For a good machine learning model, model selection is a very important part as we need to find model that aligns with our defined problem, nature of the data, complexity of problem and the desired outcomes.
- Here are the basic features of Model Selection:
  - Complexity: Consider the complexity of the problem and the nature of the data when choosing a model.
  - Decision Factors: Evaluate factors like performance, interpretability and scalability when selecting a model.
  - Experimentation: Experiment with different models to find the best fit for the problem.

Step 7: Model Training

- With the selected model the machine learning lifecycle moves to model training process. This process involves exposing model to historical data allowing it to learn patterns, relationships and dependencies within the dataset.
- Here are the basic features of Model Training:
  - Iterative Process: Train the model iteratively, adjusting parameters to minimize errors and enhance accuracy.
  - Optimization: Fine-tune model to optimize its predictive capabilities.
  - Validation: Rigorously train model to ensure accuracy to new unseen data.

Step 8: Model Evaluation and Tuning

- Evaluation Metrics: Use metrics like accuracy, precision, recall and F1 score to evaluate model performance.
- Strengths and Weaknesses: Identify the strengths and weaknesses of the model through rigorous testing.
- Iterative Improvement: Initiate model tuning to adjust hyperparameters and enhance predictive accuracy.
- Model Robustness: Iterative tuning to achieve desired levels of model robustness and reliability.

Step 8: Model Evaluation and Tuning

- Evaluation Metrics: Use metrics like accuracy, precision, recall and F1 score to evaluate model performance.
- Strengths and Weaknesses: Identify the strengths and weaknesses of the model through rigorous testing.
- Iterative Improvement: Initiate model tuning to adjust hyperparameters and enhance predictive accuracy.
- Model Robustness: Iterative tuning to achieve desired levels of model robustness and reliability.

Naive Bayesian

- The Naive Bayesian classifier is based on Bayes' theorem with the independence assumptions between predictors.
- A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets.
- Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods.

Algorithm

- Bayes theorem provides a way of calculating the posterior probability, P(c|x), from P(c), P(x), and P(x|c).
- Naive Bayes classifier assume that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors.
- This assumption is called class conditional independence.

$$\underbrace{P(c \mid x)}_{\text{Posterior Probability}} = \frac{\overbrace{P(x \mid c)}^{\text{Likelihood}}\,\overbrace{P(c)}^{\text{Class Prior Probability}}}{\underbrace{P(x)}_{\text{Predictor Prior Probability}}}$$

$$P(c \mid \mathbf{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Algorithm

- P(c|x) is the posterior probability of class (target) given predictor (attribute).
- P(c) is the prior probability of class.
- P(x|c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor.

Outlook	Temperature	Humidity	Windy	Play Golf	Class
0	Rainy	Hot	High	False	No
1	Rainy	Hot	High	True	No
2	Overcast	Hot	High	False	Yes
3	Sunny	Mild	High	False	Yes
4	Sunny	Cool	Normal	False	Yes
5	Sunny	Cool	Normal	True	No
6	Overcast	Cool	Normal	True	Yes
7	Rainy	Mild	High	False	No
8	Rainy	Cool	Normal	False	Yes
9	Sunny	Mild	Normal	False	Yes
10	Rainy	Mild	Normal	True	Yes
11	Overcast	Mild	High	True	Yes
12	Overcast	Hot	Normal	False	Yes
13	Sunny	Mild	High	True	No

The zero-frequency problem

- Add 1 to the count for every attribute value-class combination (Laplace estimator) when an attribute value (Outlook=Overcast) doesn’t occur with every class value (Play Golf=no).

Advantages of Naive Bayes Classifier

- Easy to implement and computationally efficient.
- Effective in cases with a large number of features.
- Performs well even with limited training data.
- It performs well in the presence of categorical features.
- For numerical features data is assumed to come from normal distributions

Disadvantages of Naive Bayes Classifier

- Assumes that features are independent, which may not always hold in real-world data.
- Can be influenced by irrelevant attributes.
- May assign zero probability to unseen events, leading to poor generalization.

Applications of Naive Bayes Classifier

- Spam Email Filtering: Classifies emails as spam or non-spam based on features.
- Text Classification: Used in sentiment analysis, document categorization, and topic classification.
- Medical Diagnosis: Helps in predicting the likelihood of a disease based on symptoms.
- Credit Scoring: Evaluates creditworthiness of individuals for loan approval.
- Weather Prediction: Classifies weather conditions based on various factors.