Applying the Machine Learning Workflow to KRA Datasets

To help you apply the **Machine Learning (ML) workflow** to your department’s real-world datasets, the following questions will guide you through each stage of the process. These questions will help define the problem, explore available data, preprocess and engineer features, select and train models, and evaluate performance. By systematically answering them, you will be able to implement ML solutions that align with your department’s objectives and improve data-driven decision-making.

**1. Problem Definition**

* What specific business or operational problems are you trying to solve using ML?
* How does solving this problem add value to your department’s daily operations?
* What type of machine learning task is best suited (e.g., classification, regression, clustering)?
* What are the success criteria or KPIs that will define a good model?

**2. Data Preprocessing**

* Are there missing values in your dataset? How will you handle them?
* Are there duplicate records that need to be removed?
* Are there any inconsistencies in the data, such as incorrect or unrealistic values?
* Do you need to normalize or standardize numerical features?
* Do categorical variables need encoding (e.g., one-hot encoding, label encoding)?

**3. Exploratory Data Analysis (EDA)**

* What trends or patterns can you observe in the dataset?
* Are there any strong correlations between features?
* Are there outliers that could affect model performance?
* What visualizations (histograms, box plots, scatter plots) can help you better understand the data?

**4. Feature Engineering**

* Which features contribute the most to solving the problem?
* Are there redundant or highly correlated features that can be removed?
* Can you create new features that provide better insights (e.g., aggregations, domain-specific transformations)?
* Do you need to apply dimensionality reduction techniques (e.g., PCA, feature selection)?

**5. Model Selection**

* Which machine learning algorithms are best suited for this problem (e.g. logistic regression, linear regression, SVM)?

**6. Model Training**

* What training/test split will ensure a good balance between training and validation?

**7. Model Evaluation**

* Which metrics best measure model performance (e.g., accuracy, precision, recall, F1-score, RMSE)?
* How will you interpret and explain the model's predictions?