Implementation of First Module

Exploratory Data Analysis (EDA), Label Encoding, and Standardization :

The first module mainly focused on preparing the dataset properly for machine learning. We performed **EDA**, **Label Encoding** for categorical features, and **Standardization** for numerical features.

Steps Followed:

1. **Dataset Loading:**
    - Imported the Kaggle dataset (Final_medicine_quality_dataset.csv) using pandas.
    - Checked the number of records (1100 rows) and columns (9 features).
2. **Checking Missing Values:**
    - Used isnull().sum() to find missing entries.
    - Missing values were handled appropriately:
        - For numerical columns like Days Until Expiry, missing values were filled using the mean.
3. **Univariate Analysis:**
    - Plotted histograms and boxplots for all features such as:
    - Helped understand the spread, skewness, and outliers in the data.

4. **Outlier Detection:**
    - Identified outliers using boxplots.
    - Planned to manage extreme outliers during the modeling phase.
5. **Target Variable Distribution:**
    - Plotted bar graphs for Safe vs Not Safe.
    - Found that the dataset is balanced enough for binary classification.
6. **Label Encoding:**
    - Applied Label Encoding to convert the categorical feature Active Ingredient into numerical format.
    - This was necessary because ML models cannot handle text directly.
7. **Standardization of Numerical Features:**
    - Standardized important numerical columns (like Storage Temperature, Dissolution Rate, Disintegration Time, Impurity Level, and Assay Purity).
    - Used **StandardScaler** to bring these features to a common scale (mean = 0, standard deviation = 1).
    - This improved model training speed and prediction performance.