1.Machine Learning Approach

The primary objective of this project is to develop machine learning models that extract both continuous values (e.g., weight, volume) and categorical units (e.g., kg, liters) from product descriptions and images. The approach involves using a combination of regression techniques for predicting continuous values and classification techniques for predicting categorical units. The entire process consists of the following key steps:

- **Image Processing and Text Extraction**: Preprocess images to extract textual information using Optical Character Recognition (OCR) tools like EasyOCR.
- **Feature Engineering**: Transform textual data (e.g., product descriptions) into numerical features using TF-IDF (Term Frequency-Inverse Document Frequency). Convert categorical variables (entity names) into numerical labels using label encoding.
- **Model Training and Evaluation**: Train machine learning models to predict continuous values and classify units using XGBoost and RandomForest algorithms.

2. Machine Learning Models Used

2.1. Value Prediction (Regression Model)

- **Algorithm**: RandomForest Regressor (RandomForestRegressor)
- **Features Used**:
    - **TF-IDF Vectorized Descriptions**: Text data (product descriptions) transformed into numerical vectors using TF-IDF.
    - **Encoded Entity Name**: Entity names (e.g., product type) converted into numerical values via label encoding.
- **Training Process**:
    - RandomForest Regressor is used to predict continuous values.
    - The model was trained with default hyperparameters (100 decision trees).
    - **Evaluation Metric**: Mean Absolute Error (MAE), which measures the average absolute difference between predicted and actual values.

2.2. Unit Prediction (Classification Model)

- **Algorithm**: XGBoost Classifier (XGBClassifier)
- **Objective Function**: Multi-class classification (objective='multi:softmax').
- **Features Used**:
    - **TF-IDF Vectorized Descriptions**: Same as in the regression model.
    - **Encoded Entity Name**: Similar label encoding applied to the entity names.
- **Training Process**:
    - XGBoost Classifier was trained to predict categorical units (e.g., kg, liters).

- **Evaluation Metric**: Multi-class log loss (eval_metric='mlogloss'), which measures the performance of multi-class classification.

3. Experiments

3.1. Data Preprocessing

- **Image Preprocessing**: Images were resized to 500x500 pixels and converted to grayscale to reduce complexity. EasyOCR was used to extract text from the images.
- **Textual Data Transformation**: Product descriptions were vectorized using the TF-IDF technique to represent text as sparse matrices for memory efficiency.
- **Entity Name Encoding**: Categorical entity names were label-encoded to be used as numerical features in the machine learning models.

3.2. Model Training

- For value prediction, RandomForest Regressor was trained using the transformed features.
- For unit prediction, XGBoost Classifier was trained using the same feature set to predict the product units, and its performance was evaluated using multi-class log loss and F1-score.

3.3. Memory Optimization

- Batch processing was applied during text extraction to manage memory more efficiently. Threads were used for parallelizing the extraction process.

3.4. Inference and Prediction

- Both models (regression and classification) were used to predict values and units for the test dataset. These predictions were combined into a final output.

4. Conclusion

The machine learning models were successfully trained to extract both continuous values and categorical units from product descriptions. The combined use of TF-IDF vectorization for text features and label encoding for categorical variables provided an effective representation of the data.

- **Value Prediction**: The RandomForest Regressor model performed well, with Mean Absolute Error providing an accurate measure of the prediction performance.
- **Unit Prediction**: XGBoost Classifier yielded a balanced performance with an **F1 score of 0.74**, indicating a good trade-off between precision and recall.

F1 Score = 0.74

The results demonstrate the effectiveness of combining advanced feature engineering techniques with powerful machine learning models like RandomForest and XGBoost to solve complex prediction tasks in digital marketplaces. Further optimization and hyperparameter tuning could improve the models' performance even more.