Product Matching System - Project Requirements

1. Overview

The project aims to develop a **product matching system** for a pharmaceutical marketplace. The goal is to accurately match product names from a seller's dataset to a master product list using text similarity techniques. The model should handle variations in spelling, abbreviations, and OCR errors to ensure high accuracy.

2. Objectives

- Extract relevant features from product names, including **dosage form, concentration, and price**.
- Utilize machine learning and NLP techniques to **match products accurately**.
- Ensure robustness against **spelling mistakes** and **format variations**.
- Optimize the system for CPU execution (no GPU required).

3. Data and Preprocessing

3.1 Dataset

- **Master File**: Contains official product names with unique SKUs.
- **Dataset**: Contains seller-provided product names that need matching.

3.2 Text Cleaning

- Remove **extra spaces**.
- Remove **diacritics (التشكيل)** from Arabic text.
- Convert **missing values to empty strings**.

4. Methodology

4.1 Text Similarity Computation

- Use **TF-IDF (Term Frequency-Inverse Document Frequency)** with **character-level n-grams (2 to 4 characters)**.
- Compute **cosine similarity** between **master product names** and **seller product names**.
- Assign **SKU** if the similarity score is **≥ 85%**.

4.2 Machine Learning Model for Confidence Prediction

- Train a **Random Forest classifier** to predict match reliability.
- **Input Feature**: Similarity Score (Cosine Similarity).
- **Target Variable**: 1 (Correct match), 0 (Incorrect match).
- Split data: **80% training, 20% testing**.

- Predict confidence level as **High** or **Low**.

4.3 Performance Evaluation

- **Matching Accuracy**: Percentage of correctly assigned SKUs.
- **Execution Time**: Ensure processing time is **≤ 500ms per product**.

5. Output and Deliverables

5.1 Output File (new_df.xlsx)

Contains the following columns:

- **Matched SKU**
- **Similarity Score**
- **Confidence Level** (High/Low)
- **marketplace_product_name_ar**

5.2 Terminal Output

- **Matching Accuracy (%)**
- **Average Processing Time per Record**