What Has Been Done

1. **Dataset Preparation**
    - Loaded a large maternal health dataset (4+ million rows) from a Parquet file.
    - Cleaned data: converted flags to binary, handled missing values, and ensured numeric consistency.
    - Derived the `stillbirth_risk` target from the `IS_CHILD_DEATH` column due to initial lack of labeled positives.
2. **Stratified Sampling**
    - Created a **balanced subset of 2 million rows** using:
        - All known stillbirth, maternal mortality, and child death cases.
        - Random non-critical cases to preserve class distribution.
3. **Feature Engineering**
    - Selected 43 important numerical features after excluding ID and leakage-prone columns.
    - Key features included hemoglobin metrics, anemia status, age, gravida, and parity.
4. **Modeling with Random Forest**
    - Trained a **Random Forest Classifier** with:
        - SMOTE oversampling to handle class imbalance.
        - 5-fold stratified cross-validation.
        - Class weighting (1:10) to emphasize positive class (stillbirth cases).
        - Evaluation at multiple thresholds (0.3 to 0.6) to balance recall vs precision.
5. **Evaluation**

    - Used metrics like **AUC, F1, recall, precision, and accuracy**.
    - Identified **threshold 0.6** as best tradeoff (F1: 0.0568).
    - Achieved **very high recall (0.95)** at threshold 0.3, but with **low precision** (~2.5%).

6. **SHAP Analysis**
    - Used SHAP to interpret feature influence.
    - Top features: **HEMOGLOBIN_mean, anemia_mild, AGE, inadequate_weight_gain**.
    - SHAP plots confirmed hemoglobin levels and age as dominant predictors.

Recommendations to Improve the Model

1. **Improve Precision (reduce false positives)**
    - Try other **advanced models** like:
        - **LightGBM or XGBoost** (already partially implemented).
        - **CatBoost**, which handles categorical variables better.
    - Add **class-specific sampling ratios** in SMOTE (e.g., BorderlineSMOTE or ADASYN).
    - Experiment with **ensemble methods**: combine predictions from multiple models.
2. **Feature Engineering**
    - Include **temporal features**: time since first ANC visit, time gaps between checkups.
    - Engineer interaction terms (e.g., HEMOGLOBIN_mean × AGE).
    - Use **domain knowledge** to add risk scores (e.g., anemia severity score, BMI range flags).

3. **Threshold Tuning**
    - Use **Precision-Recall Curve AUC** to optimize threshold dynamically.
    - Introduce **cost-sensitive thresholds** depending on false positive/false negative trade-offs.
4. **Data Quality & Labeling**
    - Improve labeling of `stillbirth_risk` using **additional hospital records or expert annotations**.
    - Consider **multi-label prediction** (stillbirth + maternal mortality) for more robust learning.
5. **Explainability & Validation**
    - Deploy **local SHAP explanations** for high-risk predictions to validate with clinicians.
    - Validate model performance across **geographic regions or districts** to ensure fairness.
