Explanation of High Risk Pregnancy model and Best Results

This document explains a machine learning pipeline that preprocesses data, trains a Random Forest Classifier using k-fold cross-validation, evaluates performance across multiple thresholds, and performs SHAP analysis to interpret feature importance. The dataset used is `telangana_data_with_features_and_targets (1).parquet`, and the target variable is `high_risk_pregnancy`. The code is implemented in Python using libraries such as `scikit-learn`, `pandas`, `numpy`, `lightgbm`, `shap`, and `matplotlib`.

1. Data Preprocessing

Code Explanation

The preprocessing function `prepare_data_for_targets` loads a Parquet file in batches to handle large datasets efficiently. Key steps include:

- **Loading Data**: The Parquet file is read using `pyarrow.parquet` in batches to manage memory usage.
- **Required Columns Check**: Ensures essential columns like `MOTHER_ID` and `GRAVIDA` are present.
- **Numeric Column Conversion**: Converts specified numeric columns (e.g., `AGE`, `GRAVIDA`, `HEMOGLOBIN_mean`) to numeric types, handling non-numeric values by coercing them to `NaN`.
- **Handling `GRAVIDA`**: Specifically cleans the `GRAVIDA` column by replacing `'nan'` strings with `NaN` and filling missing values with 0 after converting to numeric.
- **Flag Mapping**: Maps categorical flag columns (`IS_CHILD_DEATH`, `IS_DEFECTIVE_BIRTH`) to binary values (0 or 1) using a predefined `flag_map`.
- **Filling Missing Values**: Fills missing values in numeric columns with 0.
- **Feature Selection**: Excludes columns that could cause data leakage (e.g., aggregate risk scores, flags like `bp_risk`, `anemia_severe`) and selects numeric features for modeling.

Key Outputs

- The dataset is processed in batches, ensuring scalability.
- Non-numeric values in `GRAVIDA` (e.g., `'nan'`) are identified and handled.

- Features used for training include: `['GRAVIDA', 'PARITY', 'ABORTIONS', 'HEIGHT', 'HEMOGLOBIN_mean', 'age_adolescent', 'age_elderly', 'age_very_young', 'previous_loss', 'recurrent_loss', 'gravida_parity_ratio', 'inadequate_anc', 'irregular_anc', 'anemia_mild', 'anemia_moderate', 'anemia_severe', 'ever_severe_anemia', 'systolic_bp', 'diastolic_bp', 'hypertension', 'BMI', 'underweight', 'obese', 'normal_weight', 'depression', 'severe_depression', 'anxiety', 'severe_anxiety', 'weight_gain', 'weight_gain_per_week', 'inadequate_weight_gain']`.

2. Model Training and Cross-Validation

Code Explanation

A Random Forest Classifier is trained using 5-fold stratified cross-validation to ensure robust evaluation, especially given class imbalance in the `high_risk_pregnancy` target.

- **Train-Test Split**: The data is split into 80% training and 20% test sets, stratified by the target variable to maintain class distribution.
- **Cross-Validation Setup**: Uses `StratifiedKFold` with 5 splits, shuffling data with a fixed random seed for reproducibility.
- **Model Parameters**: The Random Forest Classifier is configured with:
  - `n_estimators=100`: 100 trees for robust predictions.
  - `max_depth=10`: Limits tree depth to prevent overfitting.
  - `min_samples_split=50`, `min_samples_leaf=25`: Ensures stability in tree splits.
  - `class_weight='balanced'`: Addresses class imbalance.
  - `random_state=42`: Ensures reproducibility.
  - `n_jobs=-1`: Uses all CPU cores for faster training.
- **Evaluation Metrics**: For each fold, the model is evaluated at four classification thresholds (0.1, 0.2, 0.3, 0.4) using:
  - F1 Score
  - Accuracy
  - Precision
  - Recall
  - Confusion Matrix
  - AUC (threshold-independent)
- **Model Selection**: The best model is selected based on the average F1 score across thresholds for each fold.

Key Outputs

- **Class Distribution** (per fold):
    - Training: ~1,225,573 negative (0) and ~54,426 positive (1) samples.
    - Validation: ~306,393 negative (0) and ~13,607 positive (1) samples.
- **Cross-Validation Results**:
    - AUC: 0.9857 ± 0.0007 (consistently high, indicating strong model performance).
    - Threshold 0.1:
        - F1: 0.2585 ± 0.0419 (low due to high recall but low precision).
        - Accuracy: 0.7540 ± 0.0564.
        - Precision: 0.1499 ± 0.0281.
        - Recall: 0.9707 ± 0.0073 (very high, capturing most positive cases).
    - Threshold 0.2:
        - F1: 0.7986 ± 0.1763 (improved due to better precision-recall balance).
        - Accuracy: 0.9732 ± 0.0289.
        - Precision: 0.7226 ± 0.2418.
        - Recall: 0.9579 ± 0.0018.
    - Threshold 0.3:
        - F1: 0.9672 ± 0.0019 (strong performance, balancing precision and recall).
        - Accuracy: 0.9972 ± 0.0002.
        - Precision: 0.9769 ± 0.0025.
        - Recall: 0.9576 ± 0.0019.
    - Threshold 0.4:
        - F1: 0.9717 ± 0.0021 (best overall, high precision and recall).
        - Accuracy: 0.9976 ± 0.0002.
        - Precision: 0.9865 ± 0.0033.
        - Recall: 0.9573 ± 0.0018.
- **Best Model**: Fold 2, with an average F1 score of 0.7960 across thresholds.

3. Test Set Evaluation

Code Explanation

The best model (from Fold 2) is evaluated on the test set using the same metrics and thresholds as in cross-validation.

- **Predictions**: Probability scores are generated, and binary predictions are made at thresholds 0.1, 0.2, 0.3, and 0.4.
- **Metrics**: AUC, F1 Score, Accuracy, Precision, Recall, and Confusion Matrix are computed.
- **Best Threshold**: Determined based on the highest F1 score on the test set.

Best Results

- **Test Set Metrics** (Best Model from Fold 2):

    - AUC: 0.9852 (comparable to cross-validation, indicating good generalization).
    - Threshold 0.1:
        - F1: 0.3096
        - Accuracy: 0.8176
        - Precision: 0.1845
        - Recall: 0.9615
        - Confusion Matrix: TN=310,694, FP=72,298, FN=654, TP=16,354
    - Threshold 0.2:
        - F1: 0.9369
        - Accuracy: 0.9945
        - Precision: 0.9183
        - Recall: 0.9561
        - Confusion Matrix: TN=381,546, FP=1,446, FN=746, TP=16,262
    - Threshold 0.3:
        - F1: 0.9659
        - Accuracy: 0.9971
        - Precision: 0.9760
        - Recall: 0.9560
        - Confusion Matrix: TN=382,592, FP=400, FN=749, TP=16,259
    - Threshold 0.4:
        - F1: 0.9691 (best)
        - Accuracy: 0.9974
        - Precision: 0.9828
        - Recall: 0.9557
        - Confusion Matrix: TN=382,708, FP=284, FN=754, TP=16,254
- **Best Threshold**: 0.4, with F1 Score: 0.9691.

4. SHAP Analysis

Code Explanation

SHAP (SHapley Additive exPlanations) is used to interpret the best model's predictions on the test set.

- **Explainer**: A `TreeExplainer` is created for the Random Forest model.
- **SHAP Values**: Computed for the positive class (index 1) to explain contributions to predicting `high_risk_pregnancy`.
- **Summary Plot**: A beeswarm plot is generated and saved as `shap_summary_plot.png`, visualizing feature impacts.
- **Feature Importance**: Calculated as the mean absolute SHAP values across test samples, sorted in descending order.

Key Outputs

- **Top Features by SHAP Importance**:
    1. `HEMOGLOBIN_mean`: 0.1896 (most influential, likely due to its strong correlation with pregnancy health).
    2. `anemia_moderate`: 0.0925 (indicating moderate anemia significantly impacts risk).
    3. `recurrent_loss`: 0.0661 (history of recurrent pregnancy loss is a key risk factor).
    4. `ABORTIONS`: 0.0434 (previous abortions contribute to risk).
    5. `gravida_parity_ratio`: 0.0244 (ratio of pregnancies to live births is relevant).
- **SHAP Summary Plot**: Saved as `shap_summary_plot.png`, showing how feature values affect predictions (e.g., higher `HEMOGLOBIN_mean` may reduce risk, while `anemia_moderate` increases it).

5. Notes on LightGBM Code

A portion of the code references a LightGBM model setup, but it appears incomplete and unused in the provided results. It includes:

- Calculation of `scale_pos_weight` for class imbalance.
- Setup for k-fold cross-validation with LightGBM, but no training or evaluation is executed.

This suggests the Random Forest model was the primary focus, and LightGBM was not fully integrated into the pipeline.

6. Conclusion

The Random Forest Classifier, trained with 5-fold cross-validation, achieves strong performance for predicting `high_risk_pregnancy`, with the best results at a threshold of 0.4 (F1=0.9691 on the test set). The model generalizes well (AUC=0.9852 on test set) and identifies key risk factors via SHAP analysis, with `HEMOGLOBIN_mean`, `anemia_moderate`, and `recurrent_loss` being the most influential features. The preprocessing pipeline effectively handles large datasets and ensures robust feature selection by excluding leakage-prone columns.