Methodology

Stage 1 – Data Familiarization

Data divided into three key groups:

- **Group 1 (Indicators and Retrospective Markers):** Includes `match_*` variables (e.g., `match_pdf_after`) and target variable (`Y`).
- **Group 2 (Clinical Diagnosis):** Variables like `preeclampsia_sum`, `pregnancy_hypertension_sum`; multiple diagnoses possible per patient.
- **Group 3 (Auxiliary Modeling Variables):** Demographics, labs, blood pressure, smoking data, medical texts, prior diagnoses; categorized by PREFIX.

Stage 2 – Exploratory Data Analysis (EDA)

- Severe imbalance (~432 positive vs. ~9500 negative cases).
- **Group 1:** Removed non-informative and empty variables (`match_measure_after`).
- **Group 2:** Removed empty columns; most patients had one diagnosis, some multiple.
- **Group 3:** Removed non-predictive variable (`Init_date`), minimal missing demographic data retained untreated (handled by model).
- **Smoking:** Extensive missing data; unrealistic "years of smoking" values set to NULL.
- **LABS:** Visual validation; extreme outliers handled (`lab_papp_a_abs_last_value` transformed using log1p); correlations managed via ElasticNet regularization.
- **Num of diag cols:** Empty columns retained for regularization; extreme outlier (112 diagnoses) set to NULL; total diagnosis count excluded due to high correlations.

Additional Clinical Conclusions

- 31.5% overlap between gestational and chronic hypertension.
- Preeclampsia most common, eclampsia rare.
- Diagnosed patients slightly older, but age ranges overlapped significantly.
- Blood pressure significantly higher among diagnosed cases.
- Smoking status showed minimal correlation.
- Literature comparison: Strong correlation for blood pressure, moderate for weight; weak correlations for age and lab tests (require nonlinear modeling).

NLP Features

- Added `clinical_length` (medical text length).
- Stratified Train/Test split (merged rare groups) prior to feature engineering.
- **Embeddings:** Encoded last diagnosis text; fallback to full text if missing expression. Used multilingual-e5-base model.
- **TF-IDF:** Calculated exclusively on Train set; binary word features created based on Mutual Information; no leakage to Test set.

Regularization and Feature Selection

Data divided into three subsets for independent selection:

- **Embeddings:** LASSO selection (no standardization).
- **Words:** Binary TF-IDF features selected via LASSO.
- **Clinical Data:** Missing data imputed (-1), standardized (StandardScaler), selected via ElasticNet.

All feature selection exclusively on Train; selections identically applied to Test to avoid leakage. Combined into unified Train/Test datasets.

Modeling

- LightGBM trained with class balancing (`class_weight={0:1,1:2}`).
- Stratified K-Fold Cross-Validation exclusively on Train for stable metrics.

Results

Evaluation on Test set (threshold 0.5):

- Precision: 95%, Recall: 63% for positive class (Y=1).
- **Business-Clinical Interpretations:**
  - Top 1%: 85% case detection, lower precision (more false positives).
  - Top 5%: Good balance (Precision: 88%, Recall: 78%). Recommended for limited budget.
  - Top 10%: Highest precision (93%), lower recall. Recommended for high precision scenarios.

Feature Importance

- Key predictors: TF-IDF-derived words, specific embeddings, diastolic blood pressure.
- Analyzed using Gain and SHAP values.

Recommendations and Future Improvements

- Model severity explicitly (multi-class or ordinal classification).
- In-depth analysis of false negatives.
- NLP fine-tuning specifically for medical corpus.
- Integrate longitudinal patient visit data.
- Enhance missing data completeness (e.g., smoking).
- Include temporal trend features (weight gain, blood pressure variability).
- Dynamic budget-based referral threshold adjustments.