Feature Engineering List - Telangana Maternal Health

Features to be Engineered from Raw Data

1. Age-Based Features

Raw Column	Engineered Feature	Calculation
AGE	(age_adolescent)	1 if AGE < 18, else 0
AGE	(age_elderly)	1 if AGE > 35, else 0
AGE	(age_very_young)	1 if AGE < 16, else 0
AGE	(age_risk_score)	age_adolescent + (age_elderly × 2) + (age_very_young × 3)
4		•

2. Obstetric History Features

Raw Column	Engineered Feature	Calculation	
GRAVIDA	multigravida	1 if GRAVIDA > 1, else 0	
PARITY	(grand_multipara)	1 if PARITY > 5, else 0	
(ABORTIONS)	(previous_loss)	1 if ABORTIONS > 0, else 0	
(ABORTIONS)	<pre>(recurrent_loss)</pre>	1 if ABORTIONS ≥ 2, else 0	
◀	•	•	•

3. ANC Visit Pattern Features

Raw Column	Engineered Feature	Calculation
(TOTAL_ANC_VISITS)	(inadequate_anc)	1 if TOTAL_ANC_VISITS < 4,
TOTAL_ANC_VISITS)		else 0
(TOTAL_ANC_VISITS)	(no anc)	1 if TOTAL_ANC_VISITS = 0,
TOTAL_ANC_VISITS)	(no_anc)	else 0
(MISSANC1FLG), (MISSANC2FLG), (MISSANC3FLG), (MISSANC4FLG)	<pre>(total_missed_visits)</pre>	Sum of all MISSANC flags
G. G		1 if total_missed_visits ≥ 2,
(MISSANC*) flags	(irregular_anc)	else 0
•	•	•

4. Anemia Classification Features

Raw Column	Engineered Feature	Calculation
(HEMOGLOBIN_mean)	anemia_mild	1 if HEMOGLOBIN ≥ 10 AND < 11, else 0
(HEMOGLOBIN_mean)	(anemia_moderate)	1 if HEMOGLOBIN ≥ 7 AND < 10, else 0
(HEMOGLOBIN_mean)	(anemia_severe)	1 if HEMOGLOBIN < 7, else 0
(HEMOGLOBIN_min)	(ever_severe_anemia)	1 if HEMOGLOBIN_min < 7, else 0
■	•	• • • • • • • • • • • • • • • • • • •

5. Blood Pressure Features

Raw Column	Engineered Feature	Calculation
BP_last	<pre>(systolic_bp)</pre>	Extract first number from "120/80" format
BP_last	diastolic_bp	Extract second number from "120/80" format
<pre>systolic_bp), diastolic_bp</pre>	hypertension	1 if systolic ≥ 140 OR diastolic ≥ 90, else 0
<pre>systolic_bp), diastolic_bp</pre>	severe_htn	1 if systolic ≥ 160 OR diastolic ≥ 110, else 0
■	•	▶

6. BMI and Nutritional Status

Raw Column	Engineered Feature	Calculation
WEIGHT_max, HEIGHT	BMI	WEIGHT / (HEIGHT/100) ²
BMI	underweight	1 if BMI < 18.5, else 0
BMI	obese	1 if BMI > 30, else 0
BMI	(normal_weight)	1 if BMI ≥ 18.5 AND ≤ 25, else 0
◀		•

7. Mental Health Features

Raw Column	Engineered Feature	Calculation
PHQ_SCORE_max	depression	1 if PHQ_SCORE ≥ 10, else 0
PHQ_SCORE_max	severe_depression	1 if PHQ_SCORE ≥ 15, else 0
GAD_SCORE_max	anxiety	1 if GAD_SCORE ≥ 10, else 0
GAD_SCORE_max	severe_anxiety	1 if GAD_SCORE ≥ 15, else 0
▲	•	•

8. Weight Change Features

Raw Column	Engineered Feature	Calculation
WEIGHT_last, WEIGHT_first	<pre>weight_gain</pre>	WEIGHT_last - WEIGHT_first
<pre>weight_gain</pre> , NO_OF_WEEKS_max	<pre>weight_gain_per_week</pre>	weight_gain / NO_OF_WEEKS_max
<pre>weight_gain_per_week</pre>	<pre>inadequate_weight_gain</pre>	1 if weight_gain_per_week < 0.2 kg, else 0
■		•

9. Birth Weight Features (from Child Data)

Raw Column	Engineered Feature	Calculation
(WEIGHT_min)	(low_birth_weight)	1 if WEIGHT_min < 2500g, else 0
(WEIGHT_min)	<pre>(very_low_birth_weight)</pre>	1 if WEIGHT_min < 1500g, else 0
(WEIGHT_mean)	<pre>(avg_birth_weight_low)</pre>	1 if WEIGHT_mean < 2500g, else 0
4		▶

10. Target Variables (Outcomes to Predict)

Raw Column(s)	Target Variable	Calculation
Multiple risk factors	(high_risk_pregnancy)	1 if ≥2 risk factors present
(DELIVERY_OUTCOME)	(stillbirth_risk)	1 if contains "stillbirth", else 0
NO_OF_WEEKS_max	<pre>(premature_birth_risk)</pre>	1 if NO_OF_WEEKS < 37, else 0
[IS_MOTHER_ALIVE]	<pre>(maternal_mortality_risk)</pre>	1 if IS_MOTHER_ALIVE = 0, else 0
[IS_DEFECTIVE_BIRTH_max]	(birth_defect_risk)	1 if IS_DEFECTIVE_BIRTH = 1, else 0
MISSANC* flags	(anc_dropout)	1 if ≥2 visits missed OR <4 total visits
4	•	▶

11. Composite Risk Scores

Raw Columns	Engineered Feature	Calculation
Age features	demographic_risk	Sum of age risk indicators
Anemia features	anemia_risk_score	mild + (moderate × 2) + (severe × 3)
All clinical features	clinical_risk_score	Sum of all clinical risk indicators
All features	<pre>overall_risk_score</pre>	Weighted sum of all risk categories
4	'	•

Data Preprocessing Steps

1. Aggregation (When Merging ANC Visits)

python

```
# Multiple ANC visits per mother need aggregation:
HEMOGLOBIN → HEMOGLOBIN_mean, HEMOGLOBIN_min, HEMOGLOBIN_max
WEIGHT → WEIGHT_first, WEIGHT_last, WEIGHT_max
BP → BP_first, BP_last
PHQ_SCORE → PHQ_SCORE_max
GAD_SCORE → GAD_SCORE_max
TWIN_PREGNANCY → TWIN_PREGNANCY_max
NO_OF_WEEKS → NO_OF_WEEKS_max
ANC ID → TOTAL ANC VISITS (count)
```

2. Missing Data Handling

- Replace HEIGHT = 0 with NaN (then handle in BMI calculation)
- Fill numeric features with 0
- Extract numeric values from string BP field
- Handle infinity values in BMI calculation

3. Data Type Optimization

- Convert all binary features to int8 (saves memory)
- Keep continuous features as float32
- Store IDs as strings

🔁 Feature Engineering Pipeline Order

- 1. First Pass: Create basic binary indicators
 - Age categories
 - Simple thresholds (anemia, hypertension)
 - Direct flags (multigravida, previous_loss)
- 2. **Second Pass**: Create derived features
 - BMI from weight and height
 - BP extraction and classification.
 - Weight gain calculations
- 3. **Third Pass**: Create composite scores
 - Risk scores combining multiple features
 - Target variables based on multiple conditions

- 4. Final Pass: Handle missing values
 - Fill NaN with 0 for model compatibility
 - Ensure all features are numeric

Notes

- **Total engineered features**: ~65 from original columns
- **Memory optimization**: Use int8 for binary features (0/1)
- **String handling**: BP field needs regex extraction
- Aggregation required: Multiple ANC visits per mother
- Critical cases: Some features specifically designed to identify maternal deaths, stillbirths
- All models use same features: But weight them differently based on prediction target