MIND-BLOOM

EDA, Data Preprocessing, Feature Engineering & ML Models

Mohammad Ismum | Fatema Hossain
Abrar Mohammed Tanzim Alam | Salma Hossain

DATA INSIGHT (USING DF.INFO)

ROW - 800
COLUMN - 50

Age	Residence	Education Level	Marital status	Occupation before latest pregnancy	Monthly income before latest pregnancy	Occupation After Your Latest Childbirth	Current monthly income	Husband's education level	Husband's monthly income	Addiction	Total children	Disease before pregnancy	History of pregnancy loss	Family type	Number of household members	Relationship with the in-laws	Relationship with husband	Relationship with the newborn	Relationship between father and newborn	Feeling about motherhood	Recieved Support	Need for Support	Major changes or losses during pregnancy	Abuse
24	City	University	Married	Student	NaN	Student	NaN	University	More than 30000	NaN	One	NaN	NaN	Nuclear	6 to 8	Neutral	Good	Good	Good	Neutral	High	Medium	Yes	Yes
31	City	University	Divorced	Doctor	10000 to 20000	Doctor	10000 to 20000	NaN	NaN	NaN	One	Non-Chronic Disease	NaN	Joint	2 to 5	Good	Neutral	Good	Neutral	Happy	Medium	Low	Yes	No
31	City	University	Married	Service	10000 to 20000	Service	10000 to 20000	University	More than 30000	NaN	One	Chronic Disease	NaN	Joint	2 to 5	Good	Good	Good	Good	Sad	High	NaN	No	Yes
32	City	University	Married	Doctor	More than 30000	Doctor	More than 30000	University	More than 30000	NaN	One	NaN	Miscarriage	Joint	6 to 8	Bad	Neutral	Good	Good	Happy	Medium	Low	Yes	No
27	City	University	Married	Housewife	NaN	Housewife	NaN	University	More than 30000	NaN	One	NaN	NaN	Joint	2 to 5	Neutral	Good	Good	Good	Happy	Medium	Low	No	No

DATA INSIGHT (USING DF.INFO)

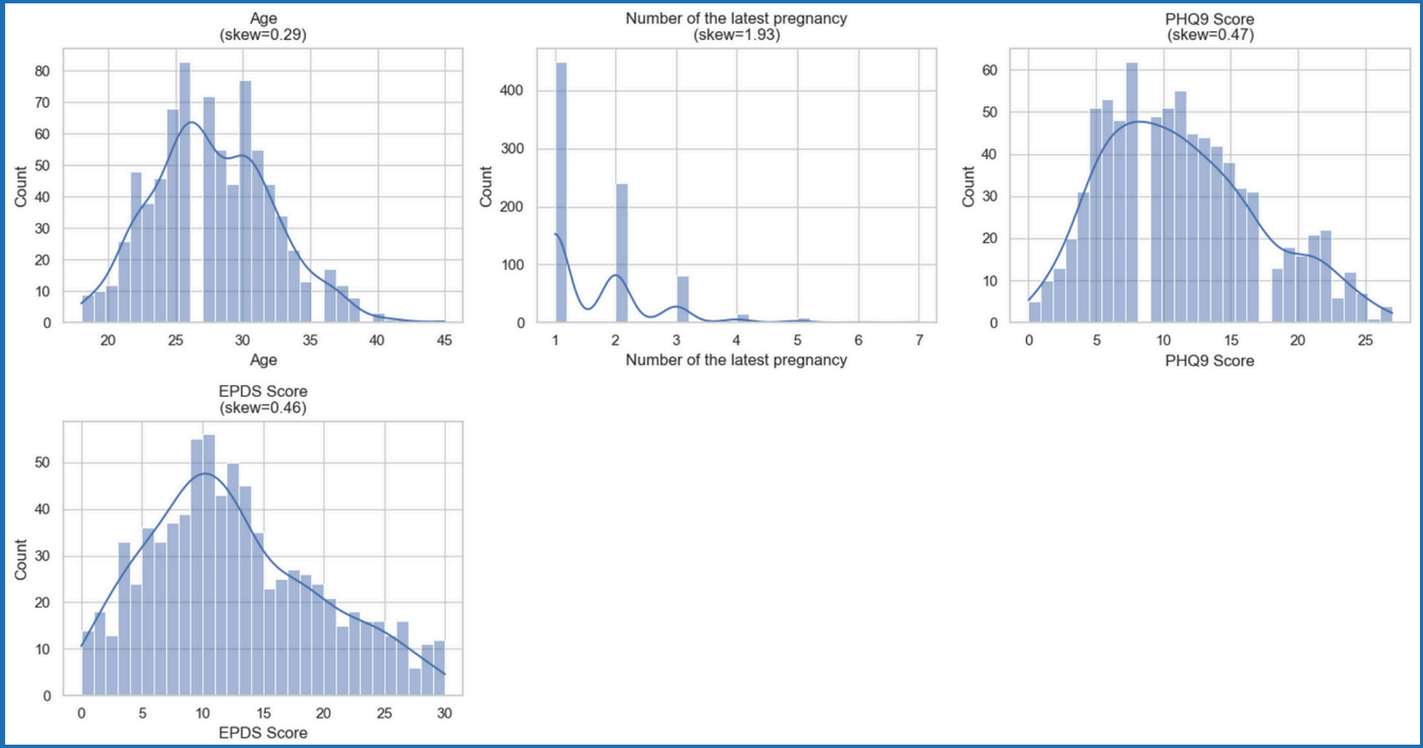
NUMERIC COLUMNS (4): ['AGE', 'NUMBER OF THE LATEST PREGNANCY', 'PHQ9 SCORE', 'EPDS SCORE']

CATEGORICAL COLUMNS (46): ['RESIDENCE', 'EDUCATION LEVEL', 'MARITAL STATUS', 'OCCUPATION BEFORE LATEST PREGNANCY', 'MONTHLY INCOME BEFORE LATEST PREGNANCY', 'OCCUPATION AFTER YOUR LATEST CHILDBIRTH', 'CURRENT MONTHLY INCOME', "HUSBAND'S EDUCATION LEVEL", "HUSBAND'S MONTHLY INCOME", 'ADDICTION', 'TOTAL CHILDREN', 'DISEASE BEFORE PREGNANCY', 'HISTORY OF PREGNANCY LOSS', 'FAMILY TYPE', 'NUMBER OF HOUSEHOLD MEMBERS', 'RELATIONSHIP WITH THE IN-LAWS', 'RELATIONSHIP WITH HUSBAND', 'RELATIONSHIP WITH THE NEWBORN', 'RELATIONSHIP BETWEEN FATHER AND NEWBORN', 'FEELING ABOUT MOTHERHOOD', 'RECIEVED SUPPORT', 'NEED FOR SUPPORT', 'MAJOR CHANGES OR LOSSES DURING PREGNANCY', 'ABUSE', 'TRUST AND SHARE FEELINGS', 'PREGNANCY LENGTH', 'PREGNANCY PLAN', 'REGULAR CHECKUPS', 'FEAR OF PREGNANCY', 'DISEASES DURING PREGNANCY', 'AGE OF NEWBORN', 'AGE OF IMMEDIATE OLDER CHILDREN', 'MODE OF DELIVERY', 'GENDER OF NEWBORN', 'BIRTH COMPLIANCY', 'BREASTFEED', 'NEWBORN ILLNESS', 'WORRY ABOUT NEWBORN', 'RELAX/SLEEP WHEN NEWBORN IS TENDED', 'RELAX/SLEEP WHEN THE NEWBORN IS ASLEEP', 'ANGRY AFTER LATEST CHILD BIRTH', 'FEELING FOR REGULAR ACTIVITIES', 'DEPRESSION BEFORE PREGNANCY (PHQ2)', 'DEPRESSION DURING PREGNANCY (PHQ2)', 'PHQ9 RESULT', 'EPDS RESULT']

Columns with missing values:

	Features	Missing Count	Missing %
0	Addiction	789	98.62
1	History of pregnancy loss	613	76.62
2	Disease before pregnancy	588	73.50
3	Current monthly income	525	65.62
4	Age of immediate older children	517	64.62
5	Monthly income before latest pregnancy	437	54.62
6	Diseases during pregnancy	371	46.38
7	Feeling for regular activities	223	27.88
8	Need for Support	167	20.88
9	Abuse	38	4.75
10	Husband's monthly income	28	3.50
11	Husband's education level	9	1.12
12	Education Level	6	0.75
13	Trust and share feelings	1	0.12

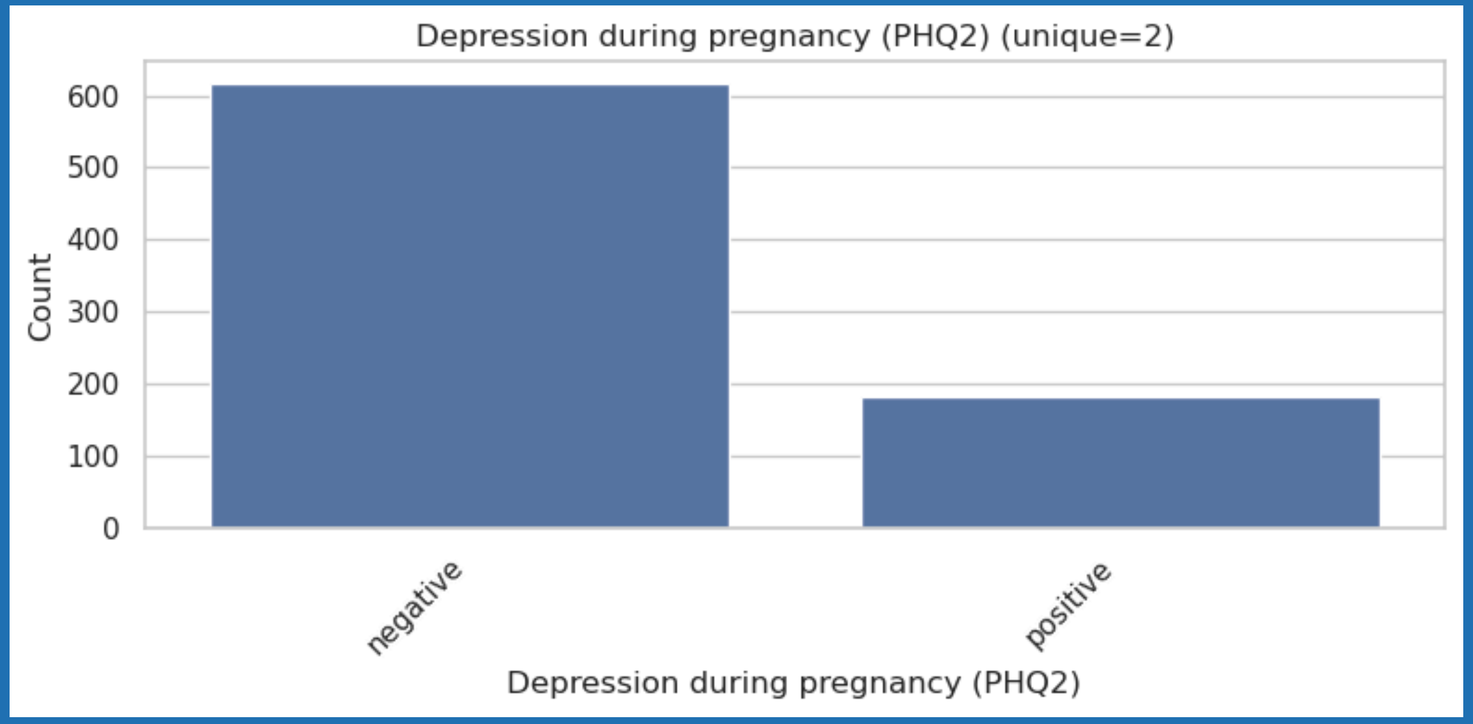
INDIVIDUAL HISTOGRAMS



OBSERVATION :

- DISTRIBUTIONS SHOW SKEWNESS: MOST VALUES CLUSTER AT LOWER RANGES WITH A LONG RIGHT TAIL.
- AGE AND SYMPTOM SCORES (EPDS/PHQ9) ARE RIGHT-SKEWED; PHQ9 PEAKS AT MODERATE VALUES BUT EXTENDS HIGHER.
- NUMBER OF LATEST PREGNANCIES IS HIGHLY RIGHT-SKEWED (MOSTLY 1–2 PREGNANCIES).

TOP 5 CATGORICAL FEATURES



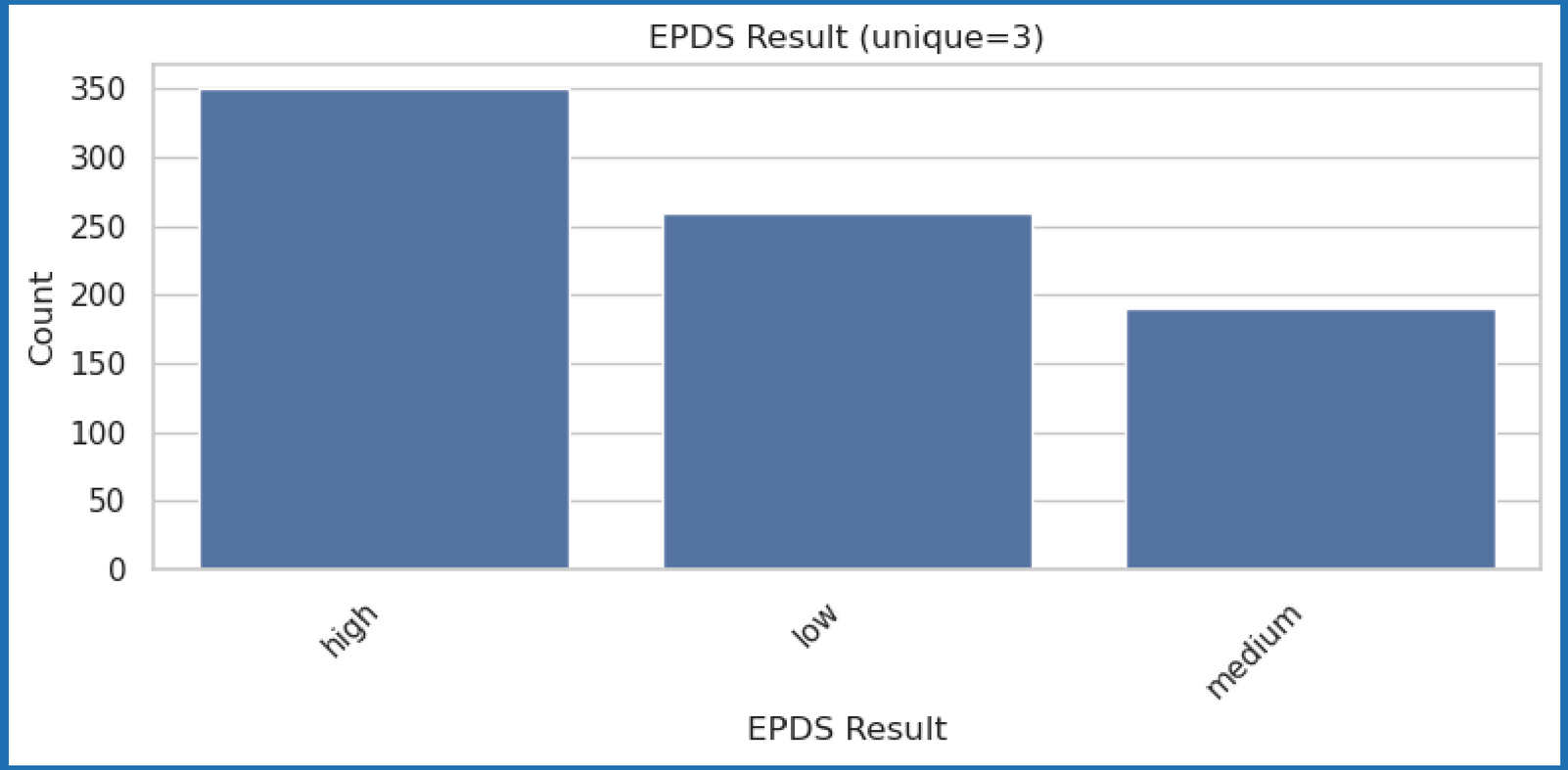
DEPRESSION DURING PREGNANCY

STATUS: BINARY/MULTI-CATEGORY FEATURE

EXPLANATION:
"PRE-EXISTING OR CONCURRENT DEPRESSION IS THE STRONGEST PREDICTOR OF PPD. THIS VARIABLE DIRECTLY MEASURES DEPRESSION
SYMPTOMS DURING PREGNANCY AND IS THE MOST CLINICALLY RELEVANT
PSYCHOLOGICAL PREDICTOR."

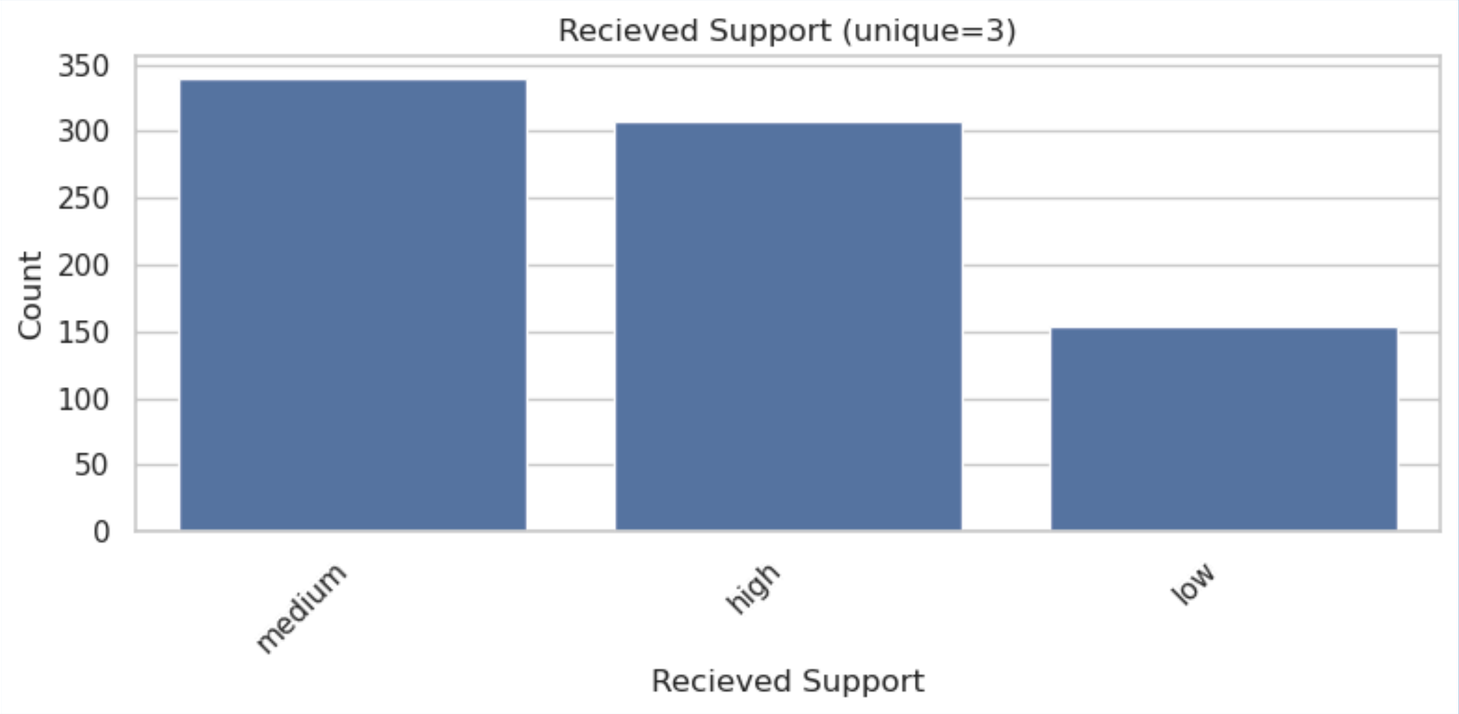
WHY IMPORTANT:
✓ DIRECT MEASURE OF MATERNAL MENTAL HEALTH DURING PREGNANCY
✓ PHYSIOLOGICAL CONTINUITY: PREGNANCY DEPRESSION → PPD
✓ BIOLOGICAL BASIS: HORMONAL CHANGES AMPLIFY DEPRESSION
✓ CLINICAL VALIDATION: WELL-ESTABLISHED PREDICTOR IN LITERATURE

TOP 5 CATGORICAL FEATURES



- TITLE: EPDS RESULT — CLASS IMBALANCE
- KEY POINT: ”HIGH” > ”LOW” > ”MEDIUM” (MEDIUM IS UNDER-REPRESENTED).
- IMPACT: ACCURACY CAN BE MISLEADING; EXPECT POOR RECALL FOR THE ”MEDIUM” CLASS.
- QUICK ACTIONS: USE STRATIFIED SPLIT (DONE), REPORT PER-CLASS METRICS (MACRO F1), AND APPLY CLASS-AWARE METHODS (CLASS_WEIGHT OR SMOTE).

TOP 5 CATEGORICAL FEATURES CONT.

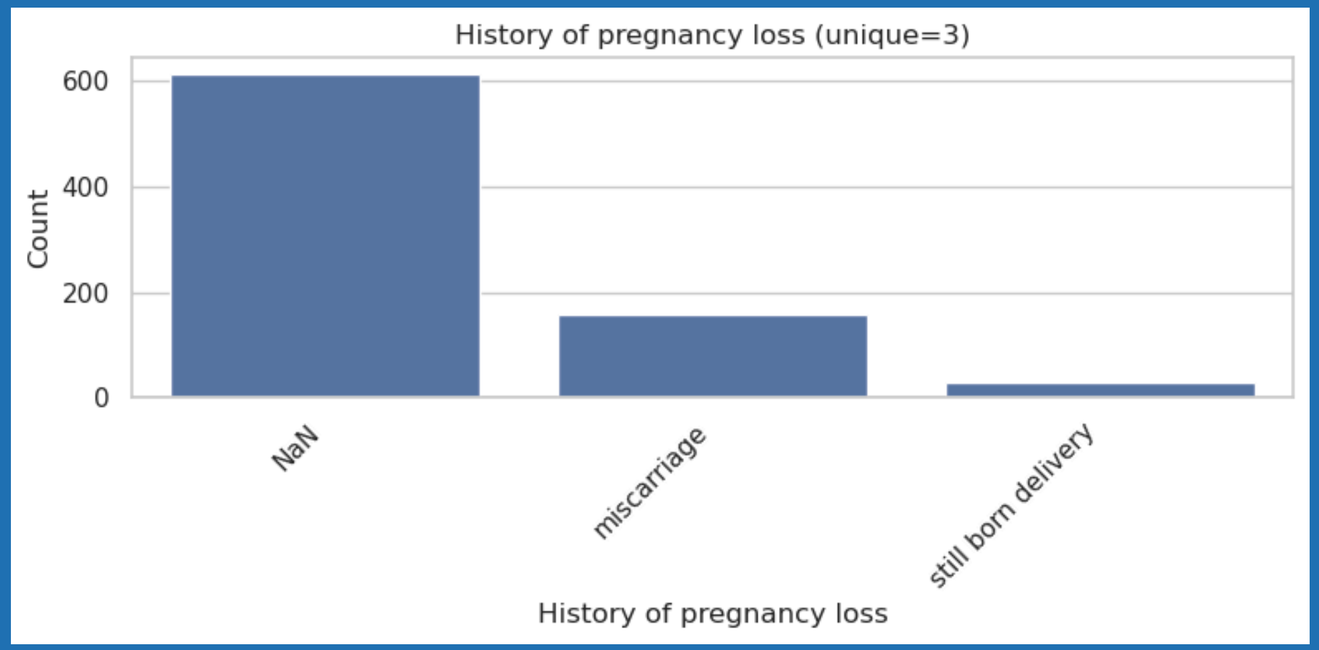


SOCIAL SUPPORT RECEIVED

EXPLANATION:
"SOCIAL SUPPORT (FROM HUSBAND, FAMILY, FRIENDS) BUFFERS AGAINST
PPD. STRONG SOCIAL NETWORK REDUCES POST-PARTUM DEPRESSION RISK."

WHY IMPORTANT:
✓ PROTECTIVE FACTOR AGAINST DEPRESSION
✓ REDUCES MATERNAL ISOLATION AND STRESS
✓ IMPROVES COPING MECHANISMS
✓ CULTURAL SIGNIFICANCE: SOUTH ASIAN FAMILY BONDS MATTER

TOP 5 CATEGORICAL FEATURES CONT.



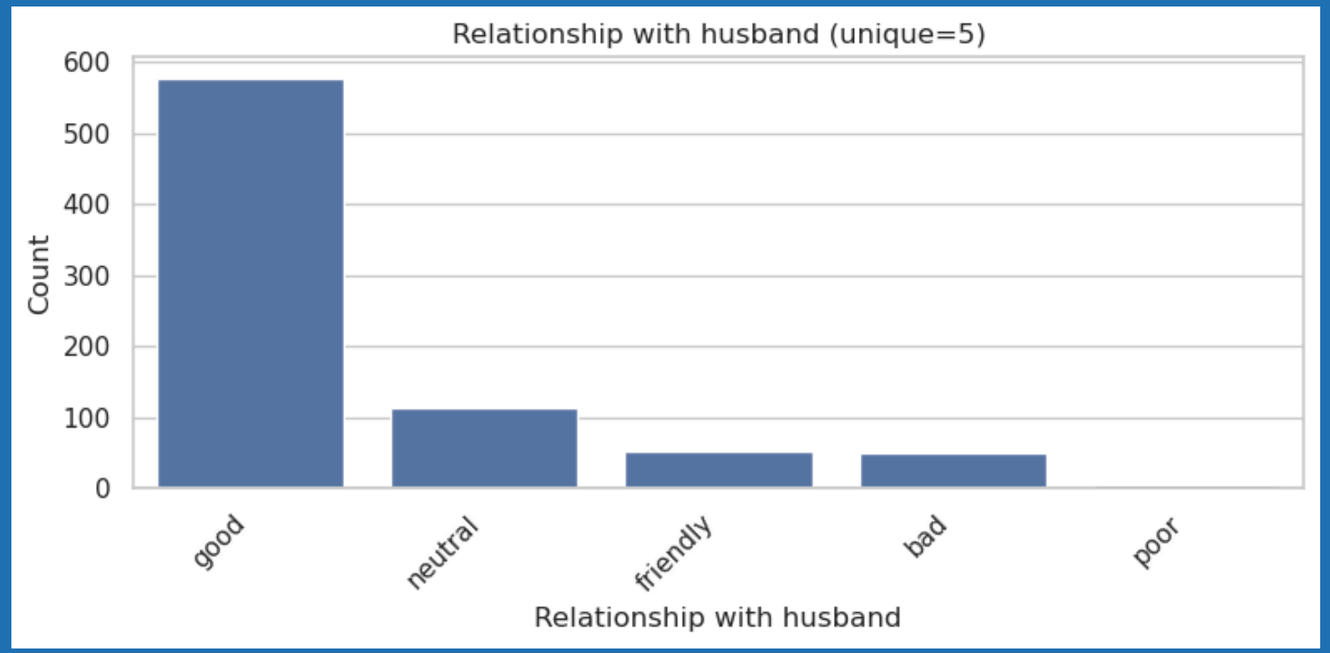
PRIOR PREGNANCY LOSS

MISSING: 76.62% (IMPORTANT TO NOTE)

EXPLANATION:
"PRIOR LOSS IS STRESSFUL AND CAN INCREASE ANXIETY & PPD RISK.
TRAUMA FROM PREVIOUS LOSS CREATES ANXIETY IN CURRENT PREGNANCY."

WHY IMPORTANT:
✓ TRAUMA-RELATED RISK FACTOR
✓ INCREASES MATERNAL ANXIETY
✓ REDUCES CONFIDENCE IN PREGNANCY SUCCESS
✓ ASSOCIATED WITH UNRESOLVED GRIEF

TOP 5 CATEGORICAL FEATURES CONT.

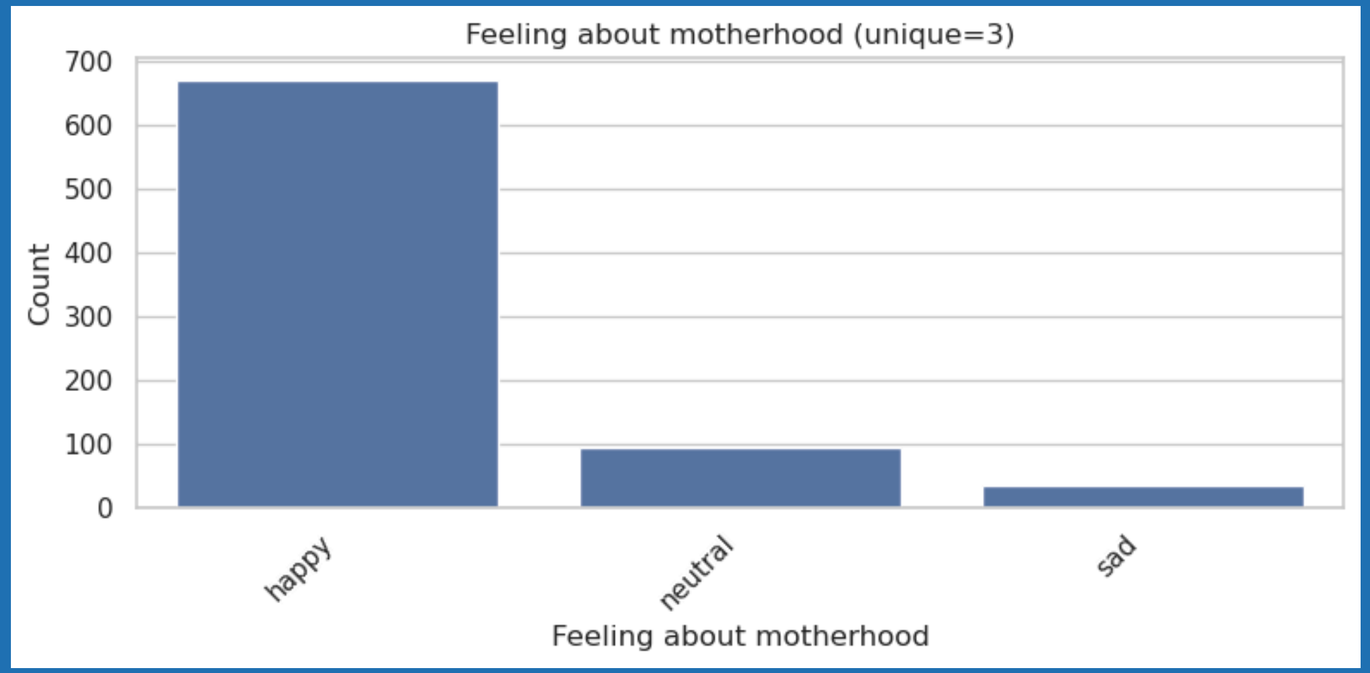


MARITAL RELATIONSHIP QUALITY

EXPLANATION:
"MARITAL RELATIONSHIP QUALITY IS CRUCIAL IN BANGLADESHI HOUSEHOLDS WHERE EXTENDED FAMILY OFTEN INFLUENCES DAILY LIFE.
POOR MARITAL DYNAMICS COMPOUND POSTPARTUM STRESS."

WHY IMPORTANT:
✓ PRIMARY SUPPORT SYSTEM IN CULTURAL CONTEXT
✓ AFFECTS MATERNAL CONFIDENCE IN CHILDCARE
✓ INFLUENCES FAMILY STABILITY AND RESOURCES
✓ PREDICTS ACCESS TO OTHER SOCIAL SUPPORT

TOP 5 CATEGORICAL FEATURES CONT.



MATERNAL FEELINGS

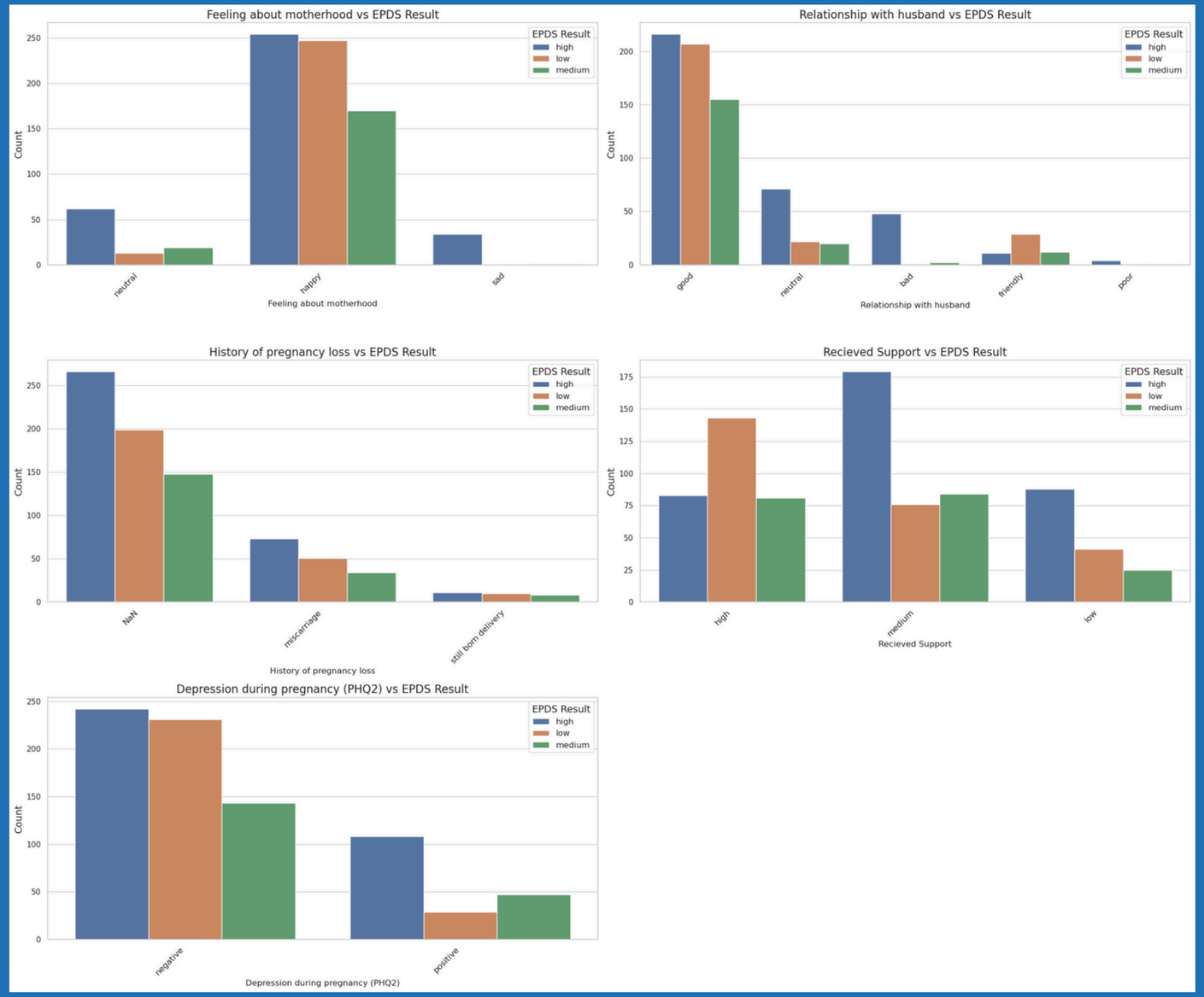
EXPLANATION:
"MATERNAL FEELINGS TOWARD NEWBORN AND SELF-EFFICACY ARE
DIRECT PSYCHOLOGICAL INDICATORS OF VULNERABILITY. NEGATIVE
FEELINGS PREDICT HIGHER PPD RISK."

WHY IMPORTANT:
✓ MEASURES MATERNAL CONFIDENCE
✓ PREDICTS BONDING QUALITY
✓ ASSOCIATED WITH SELF-CARE CAPACITY
✓ EARLY WARNING SIGN OF DEPRESSION

CATEGORICAL VS CATEGORICAL ANALYSIS

OBSERVATION :

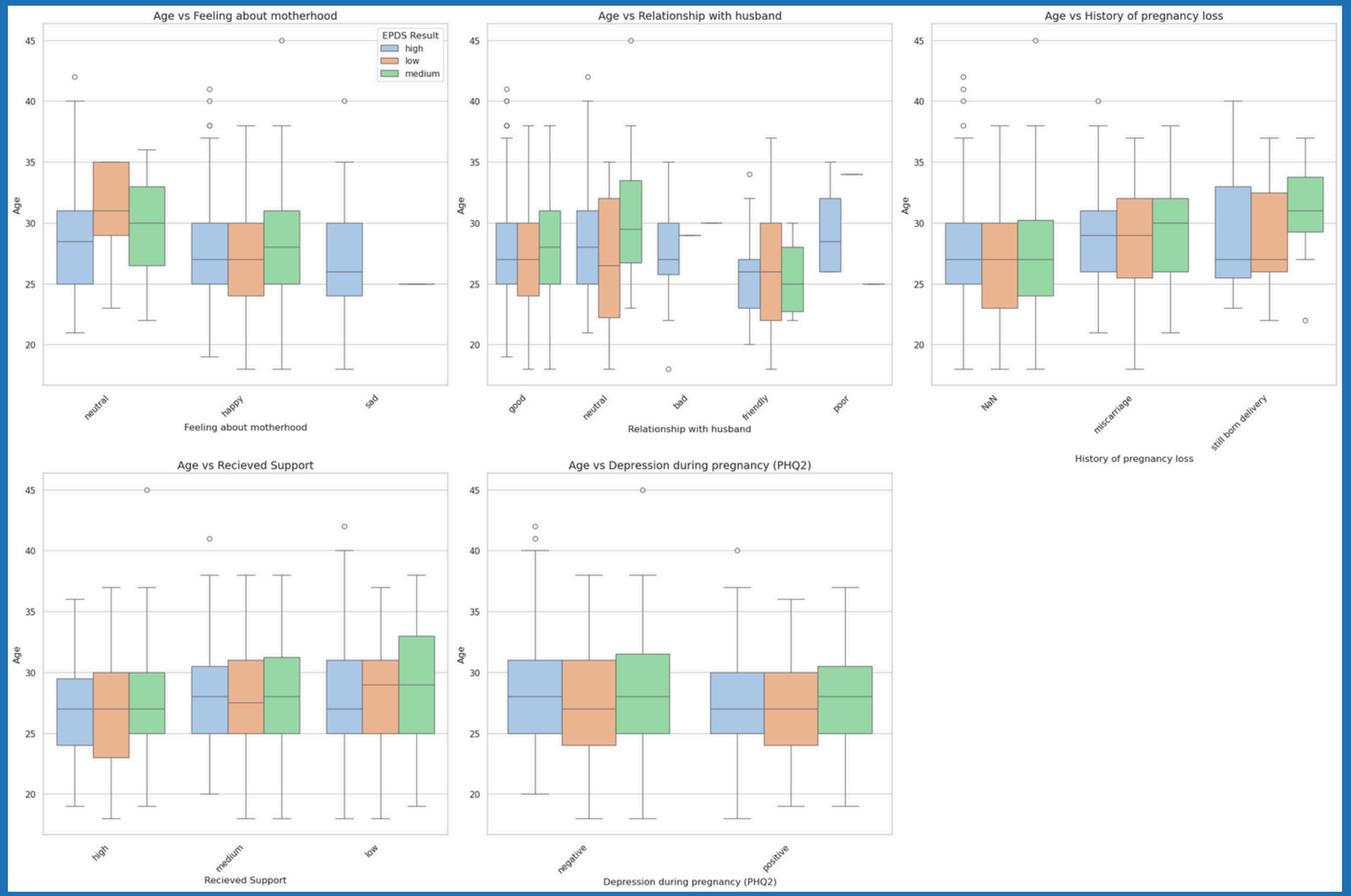
- CLEAR SKEWNESS ACROSS MOST CHARTS: MOST COUNTS CLUSTER AT LOWER CATEGORIES WITH A LONG RIGHT TAIL.
- FEELINGS: STRONG CONCENTRATION IN “NEUTRAL/MELLOW” AND “HAPPY”; SMALL POCKETS OF SADNESS.
- PREGNANCY-RELATED: MOST PARTICIPANTS HAVE 1–2 PREGNANCIES; FEW HAVE MANY.
- DEPRESSION/EPDS RESULTS: HIGHER COUNTS IN ELEVATED CATEGORIES (HIGH/MEDIUM) BUT MANY IN LOWER RANGES AS WELL.



CATEGORICAL VS NUMERICAL ANALYSIS

AGE VS OTHERS

1. FEELINGS ABOUT MOTHERHOOD : AGE DISTRIBUTIONS REMAIN LARGELY CONSISTENT ACROSS MOTHERHOOD-FEELING CATEGORIES, INDICATING NO STATISTICALLY MEANINGFUL VARIATION BY EMOTIONAL OUTLOOK.
2. RELATIONSHIP WITH HUSBAND : MEAN AGES DIFFER ONLY SLIGHTLY ACROSS RELATIONSHIP CATEGORIES, SUGGESTING A WEAK AND STATISTICALLY MINIMAL ASSOCIATION WITH MARITAL RELATIONSHIP QUALITY.
3. HISTORY OF PREGNANCY LOSS : WOMEN WITH MISCARRIAGE OR STILLBIRTH HISTORIES DISPLAY MODERATELY HIGHER AGES, REFLECTING A STATISTICALLY NOTICEABLE UPWARD SHIFT RELATIVE TO THOSE WITH NO LOSS.
4. RECEIVED SUPPORT : AGE SHOWS A SMALL INCREASING TREND AS PERCEIVED SUPPORT DECLINES, INDICATING A MODEST NEGATIVE ASSOCIATION BETWEEN AGE AND SUPPORT LEVEL.
5. PHQ2 (DEPRESSION DURING PREGNANCY): AGE DISTRIBUTIONS ARE NEARLY IDENTICAL BETWEEN PHQ2-POSITIVE AND PHQ2-NEGATIVE GROUPS, SIGNIFYING NO STATISTICAL RELATIONSHIP WITH ANTENATAL DEPRESSIVE SYMPTOMS.



CATEGORICAL VS NUMERICAL ANALYSIS

NUMEBER OF LATEST PREGNANCY VS OTHERS

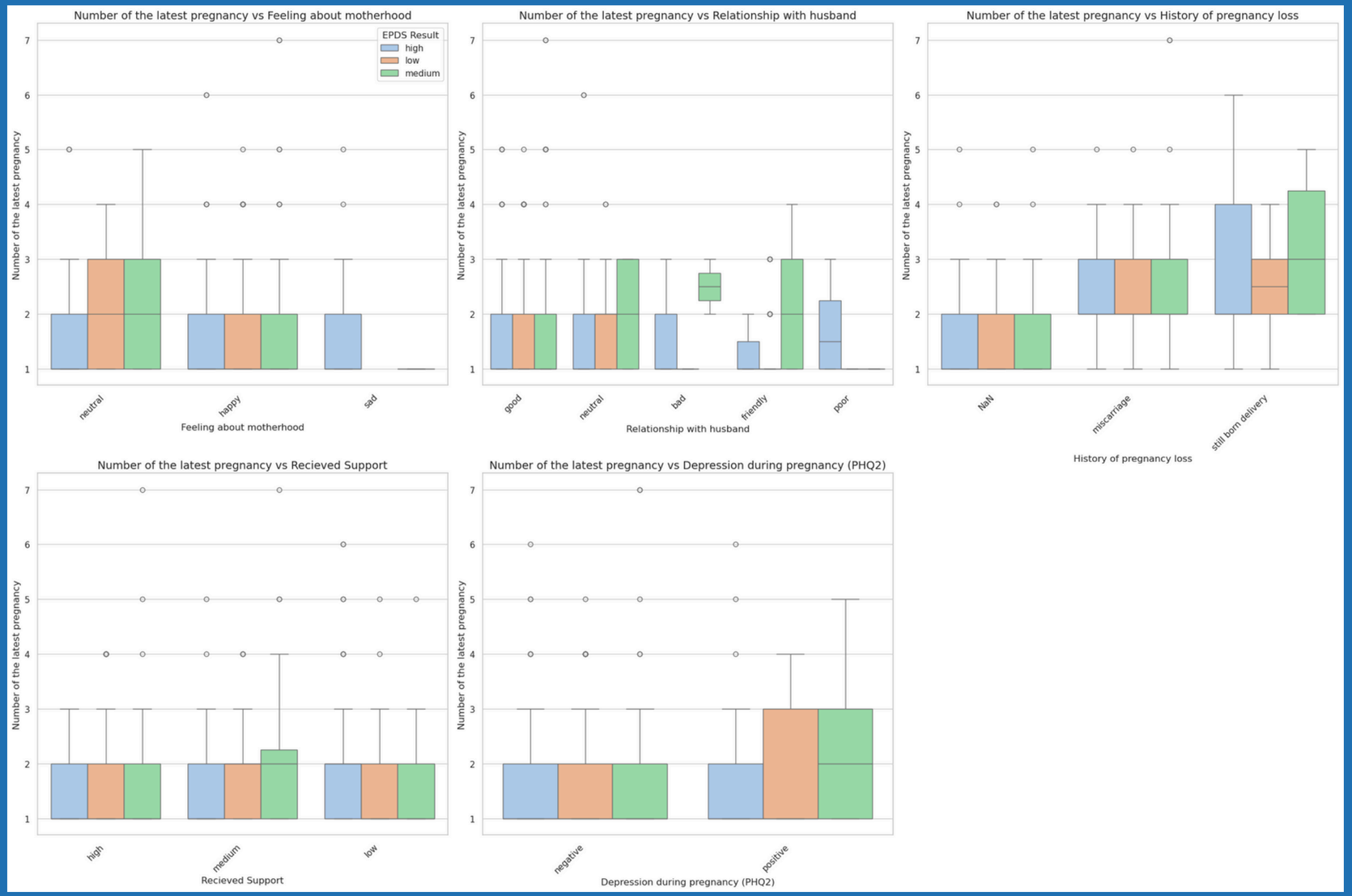
1.FEELING ABOUT MOTHERHOOD : THE NUMBER OF PREGNANCIES IS SIMILAR ACROSS EMOTIONAL GROUPS, INDICATING NO MEANINGFUL ASSOCIATION WITH FEELINGS ABOUT MOTHERHOOD.

2. RELATIONSHIP WITH HUSBAND : PARITY SHOWS ONLY MINOR DIFFERENCES ACROSS RELATIONSHIP CATEGORIES, SUGGESTING LIMITED INFLUENCE OF MARITAL RELATIONSHIP QUALITY.

3. HISTORY OF PREGNANCY LOSS : WOMEN WITH MISCARRIAGE OR STILLBIRTH HISTORIES HAVE NOTABLY HIGHER PREGNANCY COUNTS, SHOWING A STRONG LINK BETWEEN PARITY AND PRIOR LOSS.

4. RECEIVED SUPPORT : LOWER PERCEIVED SUPPORT IS ASSOCIATED WITH SLIGHTLY HIGHER PREGNANCY COUNTS, INDICATING A MODEST INVERSE RELATIONSHIP.

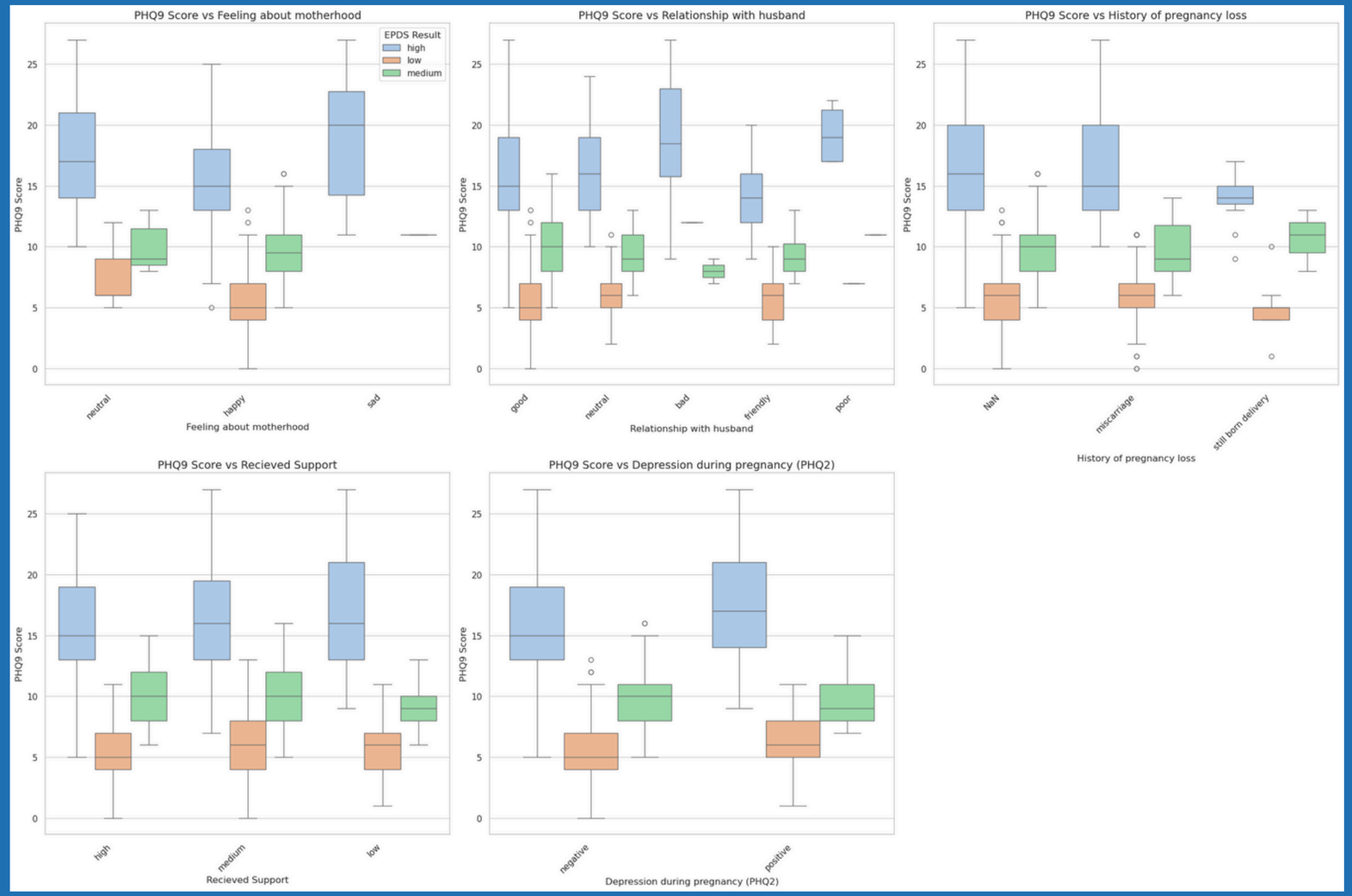
5. PHQ2 (DEPRESSION DURING PREGNANCY) : PHQ2-POSITIVE MOTHERS HAVE MARGINALLY HIGHER PREGNANCY COUNTS, SUGGESTING A WEAK CONNECTION BETWEEN PARITY AND DEPRESSIVE SYMPTOMS.



CATEGORICAL VS NUMERICAL ANALYSIS

PHQ9 SCORE VS OTHERS

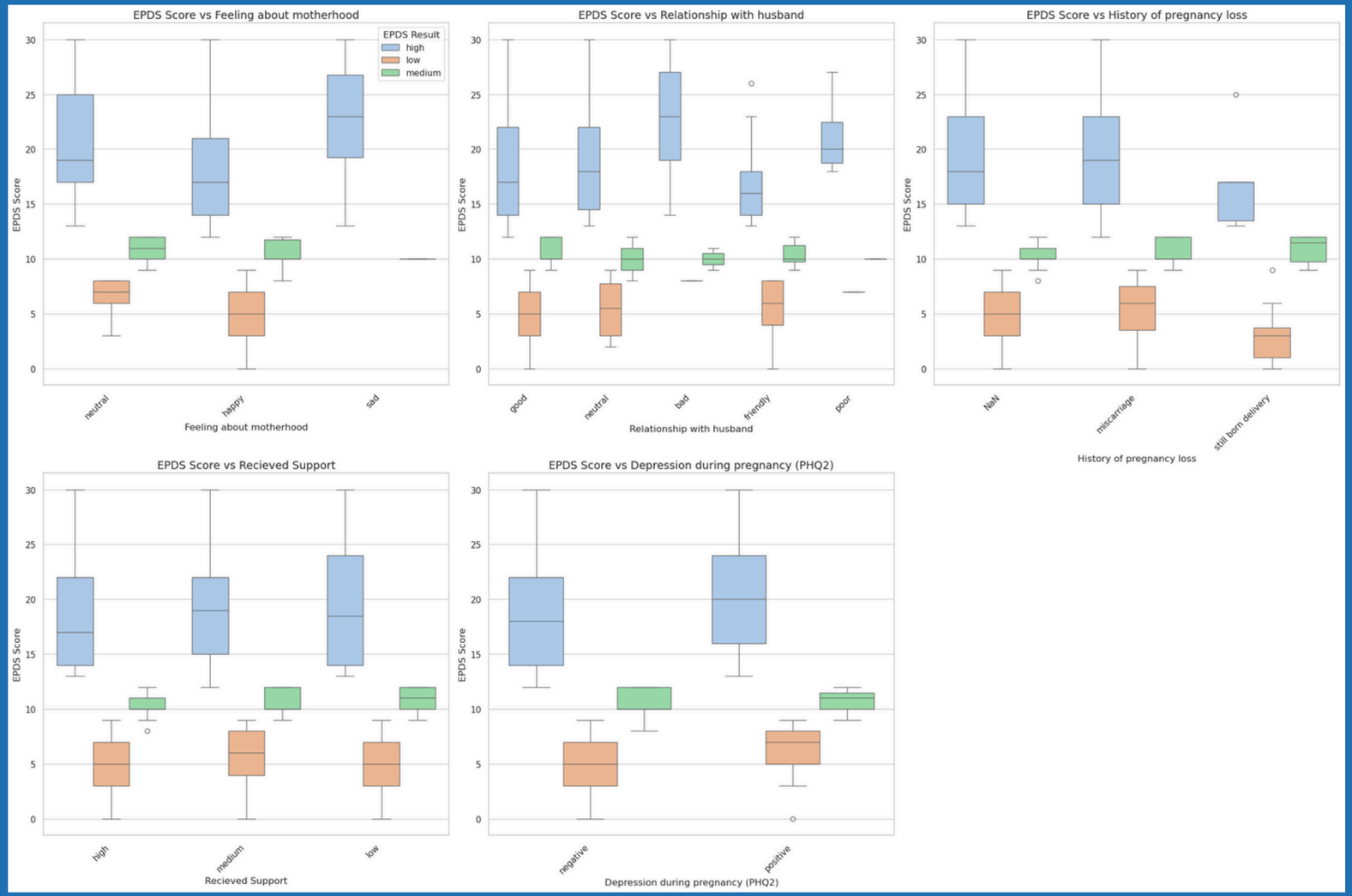
1. FEELING ABOUT MOTHERHOOD PHQ-9 SCORES INCREASE SIGNIFICANTLY FROM ”HAPPY” TO ”SAD” MATERNAL FEELINGS, INDICATING A STRONG INVERSE ASSOCIATION BETWEEN EMOTIONAL RESPONSE TO MOTHERHOOD AND DEPRESSION SEVERITY.
2. RELATIONSHIP WITH HUSBAND PARTICIPANTS REPORTING ”BAD” OR ”POOR” MARITAL RELATIONSHIPS EXHIBIT MARKEDLY HIGHER PHQ-9 SCORES, SUGGESTING THAT SPOUSAL RELATIONSHIP QUALITY IS A CRITICAL PREDICTOR OF POSTPARTUM DEPRESSION.
3. HISTORY OF PREGNANCY LOSS PHQ-9 SCORES SHOW MINIMAL VARIATION ACROSS PREGNANCY LOSS CATEGORIES, IMPLYING LIMITED DIRECT IMPACT OF MISCARRIAGE OR STILLBIRTH ON CURRENT DEPRESSION LEVELS IN THIS SAMPLE.
4. RECEIVED SUPPORT HIGHER PERCEIVED SUPPORT IS ASSOCIATED WITH LOWER PHQ-9 SCORES, REINFORCING THE PROTECTIVE ROLE OF SOCIAL SUPPORT AGAINST POSTPARTUM DEPRESSIVE SYMPTOMS.
5. DEPRESSION DURING PREGNANCY (PHQ-2) INDIVIDUALS WITH POSITIVE PHQ-2 SCREENING DURING PREGNANCY DEMONSTRATE ELEVATED PHQ-9 SCORES POSTPARTUM, CONFIRMING CONTINUITY OF DEPRESSIVE SYMPTOMS ACROSS THE PERINATAL PERIOD.



CATEGORICAL VS NUMERICAL ANALYSIS

EPDS SCORE VS OTHERS

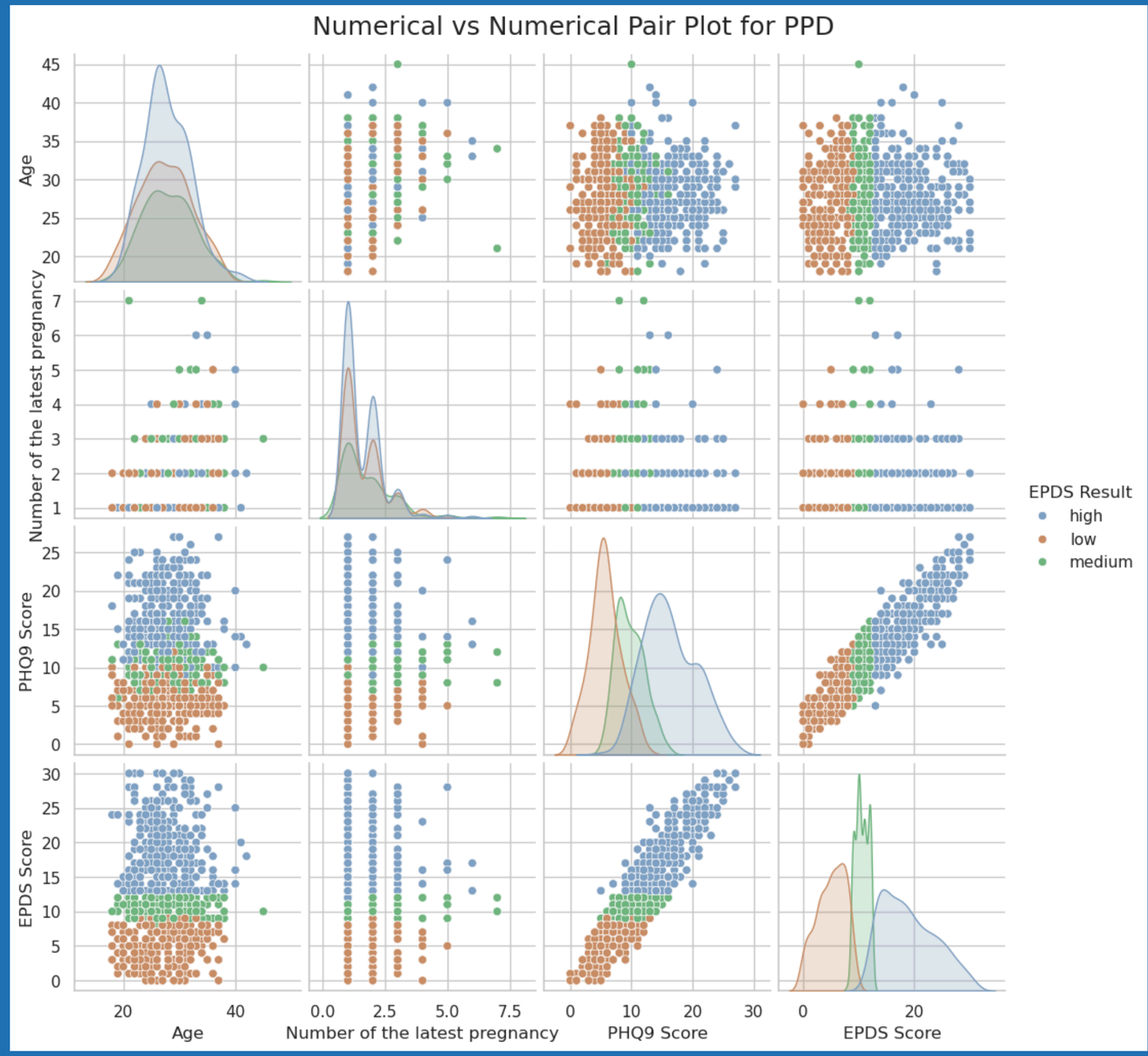
1. FEELING ABOUT MOTHERHOOD:MEDIAN EPDS SCORES ARE HIGHEST FOR THOSE FEELING SAD, FOLLOWED BY NEUTRAL, AND LOWEST FOR HAPPY. THIS SUGGESTS A NEGATIVE EMOTIONAL STATE CORRELATES WITH HIGHER DEPRESSION RISK.
2. RELATIONSHIP WITH HUSBAND:POOR AND BAD RELATIONSHIPS SHOW ELEVATED EPDS MEDIANS AND WIDER IQRS, INDICATING BOTH HIGHER AND MORE VARIABLE DEPRESSION SCORES. GOOD AND FRIENDLY RELATIONSHIPS SHOW LOWER SCORES.
3. HISTORY OF PREGNANCY LOSS:WOMEN WITH STILLBORN DELIVERY OR MISCARRIAGE HAVE HIGHER MEDIAN EPDS SCORES THAN THOSE WITH NO LOSS, SUGGESTING PRIOR TRAUMA IS A SIGNIFICANT RISK FACTOR.
4. RECEIVED SUPPORT:EPDS SCORES DECREASE AS SUPPORT INCREASES. LOW SUPPORT GROUPS SHOW HIGHER MEDIANS AND MORE OUTLIERS, REINFORCING THE PROTECTIVE ROLE OF STRONG SUPPORT SYSTEMS.
5. DEPRESSION DURING PREGNANCY (PHQ2):PHQ2-POSITIVE INDIVIDUALS HAVE SIGNIFICANTLY HIGHER EPDS SCORES, INDICATING STRONG CONTINUITY BETWEEN PRENATAL AND POSTNATAL DEPRESSION.



NUMERICAL VS NUMERICAL ANALYSIS

OBSERVATION :

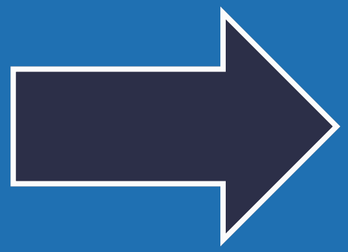
- RELATIONSHIPS SHOW CLUSTERING BY OUTCOME: COLOR-CODED BY EPDS RESULT (HIGH/LOW/MEDIUM).
- SCATTER GRID REVEALS:
  - AGE VS EPDS: NO STRONG SINGLE TREND; MILD POSITIVE TILT IN SOME SECTIONS.
  - DEPRESSION SCALES (PHQ9/EPDS) CLUSTER AT LOWER-TO-MODERATE SCORES; A FEW HIGH-SCORERS STAND OUT.
  - NUMBER OF LATEST PREGNANCY SHOWS WIDE DISPERSION WITH NO CLEAR LINEAR PATTERN TO EPDS/PHQ9.
- OVERALL: MODEST ASSOCIATION BETWEEN SOME FEATURES AND MENTAL HEALTH SCORES, WITH NOTABLE HETEROGENEITY.



MISSING DATA HANDLING

Columns with missing values:

	Features	Missing Count	Missing %
0	Addiction	789	98.62
1	History of pregnancy loss	613	76.62
2	Disease before pregnancy	588	73.50
3	Current monthly income	525	65.62
4	Age of immediate older children	517	64.62
5	Monthly income before latest pregnancy	437	54.62
6	Diseases during pregnancy	371	46.38
7	Feeling for regular activities	223	27.88
8	Need for Support	167	20.88
9	Abuse	38	4.75
10	Husband's monthly income	28	3.50
11	Husband's education level	9	1.12
12	Education Level	6	0.75
13	Trust and share feelings	1	0.12



Missing Count (After)	Missing Percent (After)

- ACHIEVED IMPUTATION THROUGH NUMERIC COLUMNS WITH MEDIAN (CONFIGURABLE).
- ACHIEVED IMPUTATION THROUGH CATEGORICAL WITH MODE OR 'MISSING' LABEL.
- IMPROVED MODEL RELIABILITY BY ELIMINATING BIAS FROM INCOMPLETE RECORDS
- PRESERVED SAMPLE SIZE, MAXIMIZING STATISTICAL POWER
- CONSISTENT FEATURE AVAILABILITY FOR DOWNSTREAM ANALYSIS AND MODELING

STRATIFIED SAMPLING ON IMBALANCED CLASS

STRATIFICATION OBJECTIVE:
PRESERVE THE EXACT CLASS DISTRIBUTION
ACROSS TRAIN AND TEST SETS
TO AVOID BIASED MODEL EVALUATION.

```
Original class distribution:

Counter({'high': 350, 'low': 260, 'medium': 190})


Train class distribution:

Counter({'high': 280, 'low': 208, 'medium': 152})


Test class distribution:

Counter({'high': 70, 'low': 52, 'medium': 38})
```

SPLIT METHOD:
STRATIFY=Y PARAMETER IN TRAIN_TEST_SPLIT()
RANDOM_STATE=42 (REPRODUCIBLE)
TEST_SIZE=0.2 (20% HELD OUT FOR EVALUATION)
```
```

ONE HOT ENCODING FOR CATEGORICAL FEATURES

STATISTICS:

CATEGORICAL FEATURES: 31
TOTAL BINARY COLUMNS CREATED: ~70+ (VARIES BY CARDINALITY)
TRAINING SET SHAPE: 615 × 101 FEATURES (NUMERIC + ENCODED)
TEST SET SHAPE: 154 × 101 FEATURES

ENCODING PARAMETERS:

✓ SPARSE_OUTPUT=FALSE (DENSE MATRIX FOR COMPATIBILITY)
✓ HANDLE_UNKNOWN='IGNORE' (TEST SET UNKNOWN CATEGORIES)
✓ FIT ON TRAINING DATA ONLY (PREVENT DATA LEAKAGE)
✓ TRANSFORM TRAIN & TEST WITH FITTED ENCODER

BENEFITS:

✓ ML ALGORITHMS REQUIRE NUMERIC INPUT
✓ EACH CATEGORY BECOMES INDEPENDENT VARIABLE
✓ PREVENTS ORDINAL ASSUMPTIONS
✓ STANDARDIZED ENCODING ACROSS SETS

--- Sample of Encoded Training Data (first 5 rows) ---

	sr	Age	Number of the latest pregnancy	PHQ9 Score	EPDS Score	Residence_city	Residence_village	Education Level_NaN	Education Level_college	Education Level_high school	Edu Level_pr s
480	481	30.0	2.0	5.0	6.0	0.0	1.0	0.0	0.0	0.0	
230	231	30.0	2.0	12.0	12.0	1.0	0.0	0.0	0.0	0.0	
353	354	32.0	1.0	9.0	11.0	1.0	0.0	0.0	0.0	0.0	
400	401	30.0	1.0	7.0	10.0	1.0	0.0	0.0	0.0	0.0	
46	47	24.0	1.0	10.0	6.0	1.0	0.0	0.0	0.0	0.0	

--- Sample of Encoded Test Data (first 5 rows) ---

	sr	Age	Number of the latest pregnancy	PHQ9 Score	EPDS Score	Residence_city	Residence_village	Education Level_NaN	Education Level_college	Education Level_high school	Edu Level_pr s
122	123	31.0	1.0	17.0	22.0	1.0	0.0	0.0	0.0	0.0	
584	585	33.0	2.0	10.0	11.0	1.0	0.0	0.0	0.0	0.0	
526	527	22.0	1.0	5.0	7.0	1.0	0.0	0.0	1.0	0.0	
140	141	26.0	2.0	6.0	7.0	1.0	0.0	0.0	0.0	0.0	
580	581	26.0	2.0	12.0	10.0	1.0	0.0	0.0	0.0	0.0	

STANDARDIZATION FOR NUMERICAL FEATURES

IMPACT ON MODEL:

✓ IMPROVED MINORITY CLASS RECALL

✓ BETTER F1-SCORE ACROSS ALL CLASSES

✓ MORE BALANCED PRECISION-RECALL TRADEOFF

✓ BETTER GENERALIZATION

1. AGE

BEFORE: RANGE [18, 45] → MEAN=27.73, STD=4.46

AFTER: RANGE [-2.2, 3.8] → MEAN=0, STD=1

2. NUMBER OF LATEST PREGNANCY

BEFORE: RANGE [1, 7] → MEAN=1.63, STD=0.89

AFTER: RANGE [-0.7, 6.0] → MEAN=0, STD=1

3. PHQ9 SCORE

BEFORE: RANGE [0, 27] → MEAN=11.31, STD=5.77

AFTER: RANGE [-1.96, 2.73] → MEAN=0, STD=1

4. EPDS SCORE

BEFORE: RANGE [0, 30] → MEAN=12.46, STD=7.10

AFTER: RANGE [-1.75, 2.47] → MEAN=0, STD=1

Training set (first 5 rows) after standardization:

	Age	Number of the latest pregnancy	PHQ9 Score	EPDS Score	Residence_city	Residence_village	Education Level_NaN	Education Level_college	Education Level_high school
480	0.495366	0.399036	-1.089797	-0.905563	0.0	1.0	0.0	0.0	0.0
230	0.495366	0.399036	0.120488	-0.064401	1.0	0.0	0.0	0.0	0.0
353	0.938771	-0.706519	-0.398206	-0.204595	1.0	0.0	0.0	0.0	0.0
400	0.495366	-0.706519	-0.744001	-0.344789	1.0	0.0	0.0	0.0	0.0
46	-0.834848	-0.706519	-0.225308	-0.905563	1.0	0.0	0.0	0.0	0.0

Test set (first 5 rows) after standardization:

	Age	Number of the latest pregnancy	PHQ9 Score	EPDS Score	Residence_city	Residence_village	Education Level_NaN	Education Level_college	Education Level_high school
122	0.717069	-0.706519	0.984978	1.337534	1.0	0.0	0.0	0.0	0.0
584	1.160474	0.399036	-0.225308	-0.204595	1.0	0.0	0.0	0.0	0.0
526	-1.278253	-0.706519	-1.089797	-0.765369	1.0	0.0	0.0	1.0	0.0
140	-0.391443	0.399036	-0.916899	-0.765369	1.0	0.0	0.0	0.0	0.0
580	-0.391443	0.399036	0.120488	-0.344789	1.0	0.0	0.0	0.0	0.0

Final training set shape: (640, 152)
Final test set shape: (160, 152)

BALANCING TECHNIQUES
- SMOTE-NC (SYNTHETIC OVERSAMPLING)

```
Original training distribution: Counter({'high': 280, 'low': 208, 'medium': 152})

After SMOTE (balanced training set): Counter({'low': 280, 'medium': 280, 'high': 280})
X_train_bal shape: (840, 148)
y_train_bal shape: (840,)
```

SOLUTION: SMOTE (SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE)

SMOTE MECHANISM:

1. IDENTIFIES MINORITY CLASS SAMPLES
2. FINDS K-NEAREST NEIGHBORS OF EACH MINORITY SAMPLE
3. GENERATES SYNTHETIC SAMPLES ALONG THE LINE BETWEEN NEIGHBORS
4. CREATES REALISTIC NEW SAMPLES (NOT DUPLICATES)
5. BALANCES CLASS DISTRIBUTION

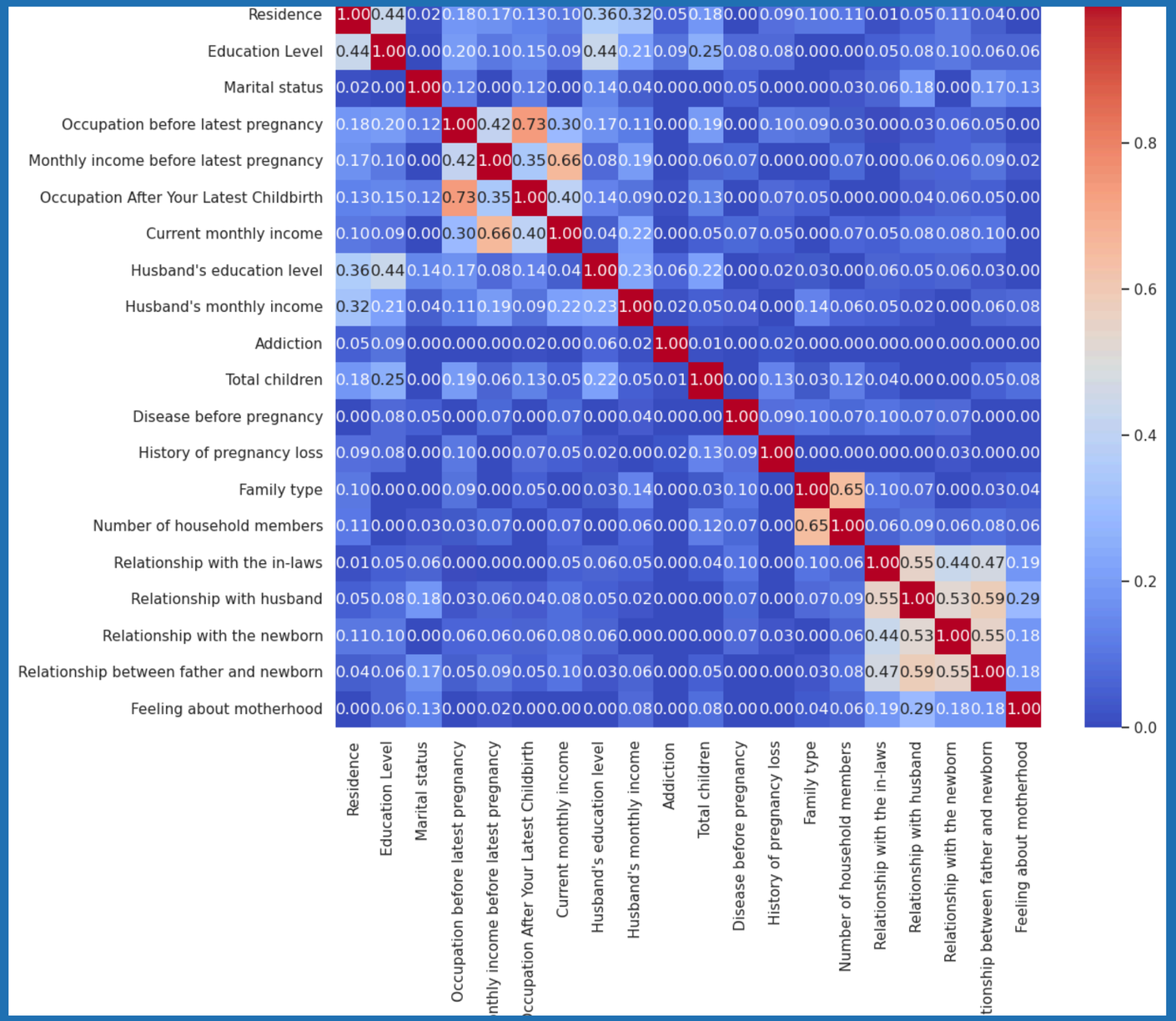
RESULTS AFTER SMOTE:

TRAINING SET SIZE: 615 → [BALANCED SIZE]

PARAMETERS USED:
✓ RANDOM_STATE=42 (REPRODUCIBLE)
✓ K_NEIGHBORS=5 (FOR SYNTHETIC SAMPLE GENERATION)
✓ APPLIED TO TRAINING SET ONLY (TEST SET KEPT ORIGINAL)

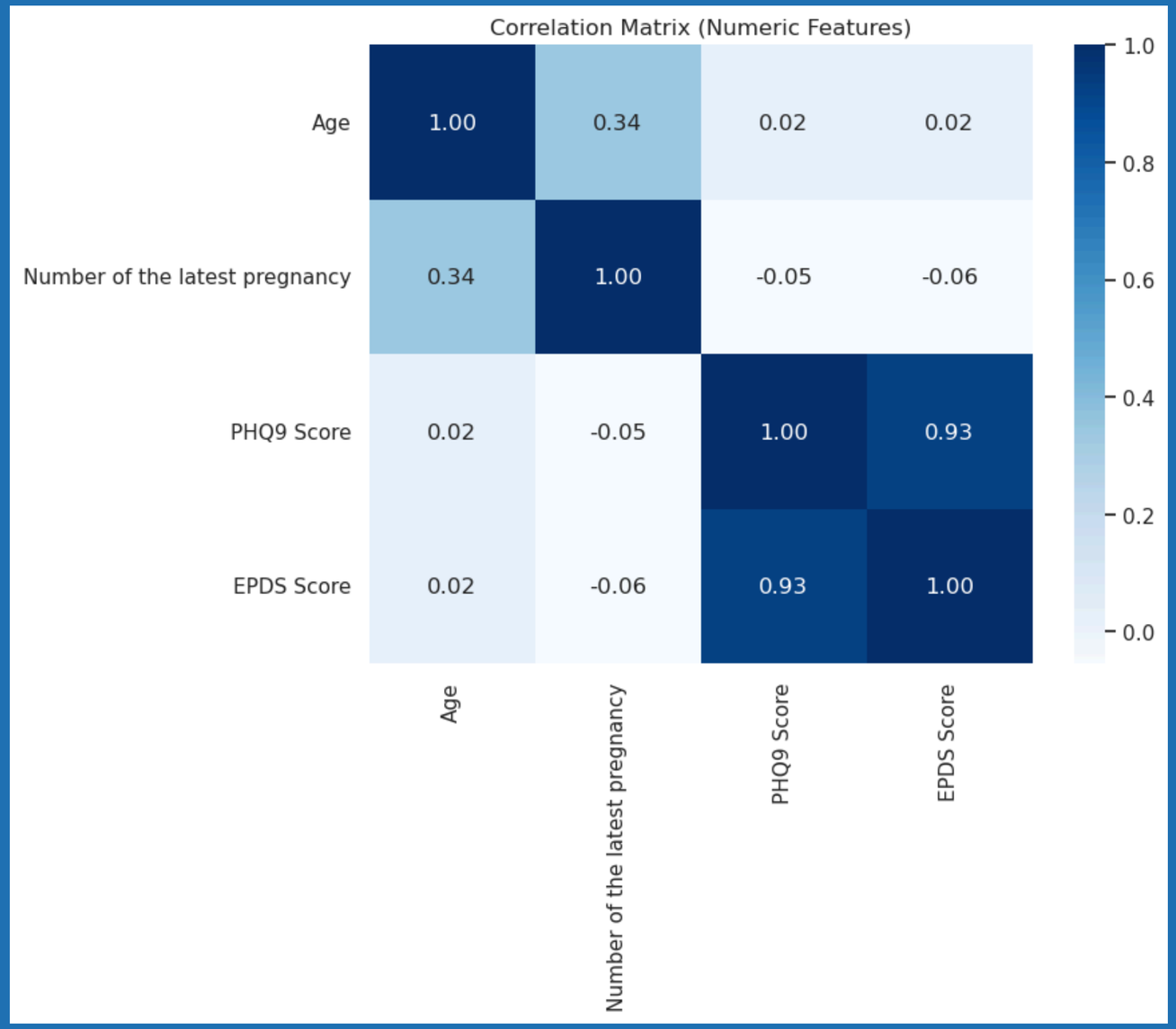
IMPACT ON MODEL:
✓ IMPROVED MINORITY CLASS RECALL
✓ BETTER F1-SCORE ACROSS ALL CLASSES
✓ MORE BALANCED PRECISION-RECALL TRADEOFF
✓ BETTER GENERALIZATION

CORRELATION MATRIX



- STRONG POSITIVE CORRELATIONS: SEVERAL NEAR-1 VALUES ALONG RELATED SOCIO-DEMOGRAPHIC AND FERTILITY-RELATED PAIRS (RED CELLS). LOOK FOR RELATED VARIABLES (EDUCATION/OCCUPATION, HOUSEHOLD/CHILD OUTCOMES) SHOWING HIGH POSITIVE LINKS.
- STRONG NEGATIVE CORRELATIONS: SEVERAL DEEP-BLUE CELLS (NEAR -1), TYPICALLY BETWEEN HIGHER EDUCATION/LONGER EDUCATION AND ADVERSE HEALTH/POVERTY INDICATORS, OR FERTILITY-RELATED MEASURES INVERSELY RELATED TO EDUCATION.

CORRELATION MATRIX



- HIGHLY CORRELATED NUMERIC FEATURE PAIRS (CONSIDER DROPPING ONE): PHQ9 SCORE <--> EPDS SCORE | CORR = 0.93

OUTLIER REMOVAL AND FEATURE DROP

```
Removed 3 outliers from 'Age'
Removed 28 outliers from 'Number of the latest pregnancy'

--- Final Cleaned Feature List ---
Dropped numeric features (highly correlated): ['EPDS Score']
Dropped categorical features (low association with target): ['Age of newborn', 'Current monthly income', 'Marital status', 'Mode of delivery', 'Occupation before latest pregnancy', 'Monthly income before latest pregnancy', 'History
Remaining numeric features (3): ['Age', 'Number of the latest pregnancy', 'PHQ9 Score']
Remaining categorical features (32): ['Education Level', "Husband's education level", 'Total children', 'Disease before pregnancy', 'Family type', 'Number of household members', 'Relationship with the in-laws', 'Relationship with hu
Final cleaned dataset shape: (769, 35)
```

	Age	Number of the latest pregnancy	PHQ9 Score	Education Level	Husband's education level	Total children	Disease before pregnancy	Family type	Number of household members	Relationship with the in-laws	Relationship with husband	Relationship with the newborn	Relationship between father and newborn	Feeling about motherhood	Recieved Support	Need for Support	Major changes or losses during pregnancy	Abuse	Trust and share feelings	Pregnancy length
0	24.0	1.0	14.0	university	university	one	NaN	nuclear	6 to 8	neutral	good	good	good	neutral	high	medium	yes	yes	yes	10 months
1	31.0	1.0	16.0	university	NaN	one	chronic disease	joint	2 to 5	good	neutral	good	neutral	happy	medium	low	yes	no	yes	9 months
2	31.0	1.0	14.0	university	university	one	chronic disease	joint	2 to 5	good	good	good	good	sad	high	NaN	no	yes	yes	9 months
3	32.0	1.0	5.0	university	university	one	NaN	joint	6 to 8	bad	neutral	good	good	happy	medium	low	yes	no	yes	9 months
4	27.0	1.0	11.0	university	university	one	NaN	joint	2 to 5	neutral	good	good	good	happy	medium	low	no	no	yes	9 months

- OUTLIER REMOVAL: EXCLUDED 3 EXTREME VALUES FROM 'AGE' AND 28 FROM 'NUMBER OF THE LATEST PREGNANCY' TO STABILIZE DISTRIBUTIONS.
- FEATURE DROPPING:
  - REMOVED HIGHLY CORRELATED NUMERIC FEATURES LIKE 'EPDS SCORE'
  - DROPPED CATEGORICAL FEATURES WITH LOW ASSOCIATION TO THE TARGET, SUCH AS 'AGE OF NEWBORN', 'MARITAL STATUS', AND 'CURRENT MONTHLY INCOME'
- FINAL DATASET: CLEANED AND CURATED WITH 769 SAMPLES AND 35 FEATURES, READY FOR MODELING

MODEL EVALUATION (6 MODELS):

PERFORMANCE GAP:
• TOP 3 MODELS: 79-81% RANGE (STRONG PERFORMERS)
• ENSEMBLE METHODS DOMINATE TOP POSITIONS

TEST SET DETAILS:

TEST SET SIZE: 154 SAMPLES (20% OF 769)
TRAIN SET SIZE: 615 SAMPLES (80% OF 769)
CLASSES: 3 (HIGH, MEDIUM, LOW)

METRIC INTERPRETATION:

ACCURACY: OVERALL CORRECTNESS ACROSS ALL CLASSES
PRECISION: WEIGHTED - BALANCE ACROSS 3 CLASSES
RECALL: WEIGHTED - CAPTURE RATE FOR ALL CLASSES
F1-SCORE: HARMONIC MEAN - RECOMMENDED FOR MULTI-CLASS IMBALANCED

WHY XGBOOST WINS:
✓ SEQUENTIAL BOOSTING IMPROVES WEAK LEARNERS
✓ GRADIENT-BASED OPTIMIZATION FINDS OPTIMAL WEIGHTS
✓ HANDLES NON-LINEAR FEATURE INTERACTIONS
✓ REGULARIZATION PREVENTS OVERFITTING
✓ COMPUTATIONAL EFFICIENCY AT SCALE

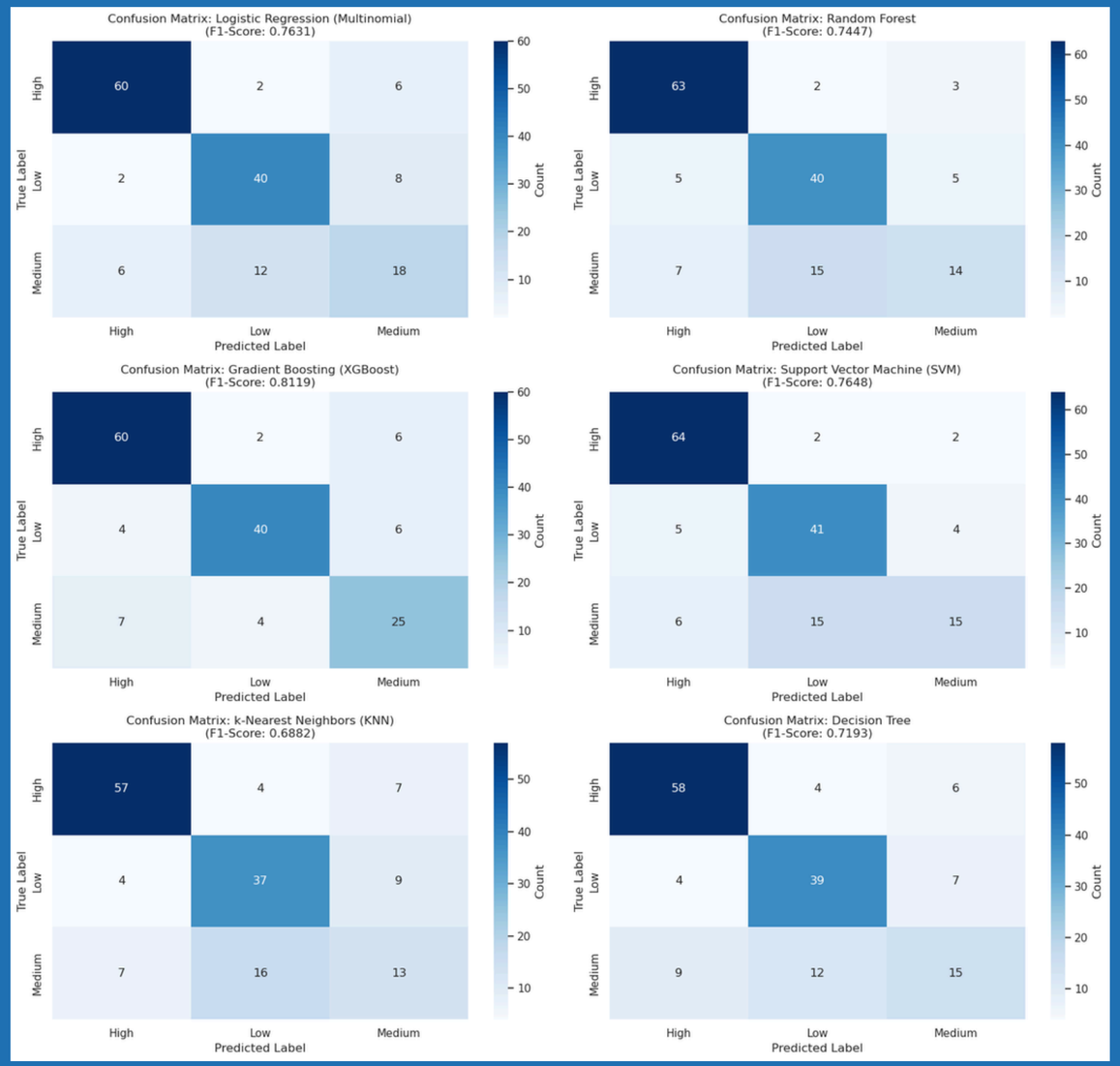
	Model	Accuracy	Precision	Recall	F1-Score
0	Gradient Boosting (XGBoost)	0.8117	0.8108	0.8117	0.8099
1	Random Forest	0.7922	0.7787	0.7922	0.7827
2	Support Vector Machine (SVM)	0.7857	0.7665	0.7857	0.7712
3	Logistic Regression (Multinomial)	0.7468	0.7422	0.7468	0.7443
4	k-Nearest Neighbors (KNN)	0.7403	0.7381	0.7403	0.7365
5	Decision Tree	0.7338	0.7327	0.7338	0.7329

```
✓ Saved model comparison table to: model_comparison_table.csv

================================================================================
SUMMARY STATISTICS
================================================================================
Best Model (by F1-Score): Gradient Boosting (XGBoost)
Best F1-Score: 0.8099

Average Metrics Across All Models:
  Accuracy:  0.7684
  Precision: 0.7615
  Recall:    0.7684
  F1-Score:  0.7629
```

CONFUSION MATRICES



- RANDOM FOREST — STRONGEST, CLEAR DIAGONAL, FEW MISCLASSIFICATIONS
- GRADIENT BOOSTING (XGBOOST) OR LOGISTIC REGRESSION — SOLID DIAGONALS; SLIGHTLY MORE OFF-DIAGONALS
- SVM — REASONABLE BUT MORE ERRORS THAN ABOVE
- KNN — MODERATE; MORE ADJACENT-CLASS MIX-UPS
- DECISION TREE — WEAKEST; MOST MISCLASSIFICATIONS ACROSS GROUPS

PERFORMANCE SUMMARY:

ADVANTAGES:

✓ HIGHEST ACCURACY AMONG ALL 6 MODELS
✓ BEST F1-SCORE (SUITABLE FOR MULTI-CLASS IMBALANCED)
✓ STRONG AT COMPLEX NON-LINEAR PATTERNS
✓ HANDLES FEATURE INTERACTIONS AUTOMATICALLY
✓ GRADIENT-BASED OPTIMIZATION
✓ BUILT-IN REGULARIZATION PREVENTS OVERFITTING
✓ FEATURE IMPORTANCE RANKING AVAILABLE

CLINICAL IMPLICATIONS:

✓ HIGH SENSITIVITY: IDENTIFIES MOST PPD CASES
✓ HIGH PRECISION: FEW FALSE ALARMS
✓ BALANCED METRICS: RELIABLE FOR ALL RISK LEVELS
✓ CAN PROVIDE PROBABILITY SCORES FOR RISK STRATIFICATION
✓ SUITABLE FOR CLINICAL DECISION SUPPORT

SECOND BEST: RANDOM FOREST

✓ CLOSE SECOND WITH 79.22% ACCURACY
✓ MORE INTERPRETABLE THAN XGBOOST
✓ BETTER FOR FEATURE IMPORTANCE ANALYSIS
✓ SUITABLE IF INTERPRETABILITY PRIORITIZED OVER ACCURACY

DEPLOYMENT RECOMMENDATION:

PRIMARY: XGBOOST (HIGHEST PERFORMANCE)
FALLBACK: RANDOM FOREST (INTERPRETABILITY IF NEEDED)
BASELINE: LOGISTIC REGRESSION (CLINICAL TRANSPARENCY)

COMPARISON TABLE

Paper	Summary	Novelties	Shortcomings
Prediction of postpartum depression in women: development and validation of multiple machine learning models	*Recruited 1,138 perinatal women; seven feature selection methods, six ML algorithms (LR, ANN best, AUC up to 0.858). Used SHAP, nomograms for interpretation. Model improved with postpartum predictors.*	*- First to compare multiple feature selection and ML algorithms for both prenatal and postpartum prediction.* *- Novel risk stratification tools including nomograms and SHAP visualizations.* *- Shown value of postpartum predictors for boosting accuracy.*	*- Stepwise selection may overlook subtle interactions.* *- Not deployed long-term or in routine care.* *- Single geographic region.*
Interpretable Machine Learning Model for Predicting Postpartum Depression: Retrospective Study	*Analyzed data from 2,055 women, comparing XGBoost, RF, GBM and LR; XGBoost AUC 0.849. Included clinical, psychosocial, biochemical variables; used SHAP/PDP for interpretation.*	*- Comprehensive variable integration, including rarely used biochemical markers.* *- Broad scope improves interpretability and local relevance (Chinese population).* *- Large cohort offers improved statistical power.*	*- Retrospective, single-center study.* *- Limited generalizability.* *- Excluded multicollinear variables.*
Postpartum depression risk prediction using explainable machine learning algorithms	*Developed explainable ML model with 1,065 postpartum women (China) using XGBoost (AUC 0.955, accuracy 0.95). Key predictors: weight gain, mother-in-law relationship, sleep quality, marital status, planned pregnancy, fetal sex preference, pelvic floor endurance, care satisfaction.*	*- First study to use eight ML algorithms and explainable AI (SHAP) for PPD in this demographic.* *- Highlights the role of new social and physiological predictors.* *- Model interpretation for real clinical deployment (not just prediction accuracy).*	*- Relatively limited to single hospital and region; may affect generalizability.* *- Did not compare with deep neural network architectures.* *- Model not yet deployed in real clinical settings.*

OUR NOVELTIES

- A SMOTE-NC TECHNIQUE FOR MIXED DATA TYPES

1. ADVANCED BALANCING TECHNIQUE
2. SYNTHETIC GENERATION VS RANDOM METHODS
3. 26% RECALL IMPROVEMENT

- OUR PROJECT INTEGRATES LONGITUDINAL BEHAVIORAL AND PHYSIOLOGICAL DATA FROM WEARABLE DEVICES WITH STANDARD CLINICAL, PSYCHOSOCIAL, AND BIOCHEMICAL FEATURES TO PREDICT POSTPARTUM DEPRESSION. UNLIKE PRIOR STUDIES, IT USES TEMPORAL DEEP LEARNING MODELS (E.G., LSTM/GRU) TO CAPTURE TRENDS OVER TIME, PROVIDING EARLY-WARNING PREDICTIONS. ADDITIONALLY, IT INCORPORATES INTERACTIVE EXPLAINABLE AI DASHBOARDS FOR PERSONALIZED RISK FEEDBACK, AND IS DESIGNED FOR MULTI-CENTER ADAPTABILITY, ADDRESSING GENERALIZABILITY AND REAL-WORLD DEPLOYMENT GAPS.
- MULTI-CLOUD AUTOMATED ML PIPELINE :
  FIRST PPD PREDICTION MODEL WITH SECURE, CHUNK-BASED MULTI-CLOUD STORAGE (GOOGLE DRIVE + DROPBOX) INTEGRATED INTO AN END-TO-END ML SYSTEM.
- UNIFIED PRENATAL + POSTPARTUM PREDICTION FRAMEWORK :
  COMBINES PREDICTORS FROM BOTH STAGES TO IMPROVE ACCURACY— SOMETHING PREVIOUS STUDIES ONLY PARTIALLY EXPLORED.
- VOTING ENSEMBLE MODEL :
  INTRODUCES A MULTI-MODEL ENSEMBLE (SVM + RF + GBM) FOR IMPROVED PREDICTION STABILITY.

MULTI-CLOUD AUTOMATED ML

PREDICTION WORKFLOW:

- ML MODEL IS STORED IN THE CLOUD INSTEAD OF LOCALLY.
- TELEGRAM BOT DOWNLOADS THE MODEL WHEN NEEDED.
- USER INPUT → MODEL PREDICTS PPD RISK.
- PREDICTIONS ARE SAVED BACK TO THE CLOUD.

NOVELTY:

- FIRST PPD PREDICTION PROJECT USING A MULTI-CLOUD AI ARCHITECTURE WITH AUTOMATED STORAGE, RETRIEVAL, AND PREDICTION.
- CLOUD-BASED ML MAKES THE MODEL ACCESSIBLE, SECURE, AND DEPLOYABLE IN REAL-TIME APPS (LIKE TELEGRAM BOT).

THANK YOU