Islington Journal of Multidisciplinary Research Vol. 1, No. 1 | pp. 12–22

Islington College, Kathmandu, Nepal • ISSN 3149-7209 (Online) Year: 2026

Original Research Article
DOI: 10.67556/hq2s5x30

Machine Learning Based Postpartum Depression Risk Prediction: A Case Study on a Bangladeshi Dataset with Transferability Discussions for Nepal

Aditi Adhikari[1,*], Nirupan Karki[2],

[1]Islington College, Kathmandu, Nepal
[2]Pulchowk Campus, Institute of Engineering, Tribhuvan University, Kathmandu, Nepal
[*]*Correspondence: adhikarisharmaaditi@gmail.com*

ARTICLE HISTORY

Received: 21 March 2026 Revised: 24 April 2026
Accepted: 23 May 2026 Published: 08 June 2026

Scan to access

How to Cite (Harvard): Adhikari, A. and Karki, N. (2026). 'Machine learning based postpartum depression risk prediction: A case study on a Bangladeshi dataset with transferability discussions for Nepal', *Islington Journal of Multidisciplinary Research*, 1(1), pp. 12–22. Available at: https://doi.org/10.67556/hq2s5x30

ABSTRACT

Postpartum depression (PPD) is a common perinatal mood disorder that often goes undetected, affecting roughly 10-15 percent of mothers, with higher rates across the Middle East and Asia and about half of cases never reported. Awareness remains low in countries such as Nepal, making early detection difficult. This study used the dataset Data for Postpartum Depression Prediction in Bangladesh (Raisa and Kaiser 2025) to test whether supervised machine learning can sort mothers into three EPDS-based risk categories (Low, Medium, High), and how much of that signal depends on the dataset's own depression-screening features rather than on sociodemographic and psychosocial variables. Four classifiers (Logistic Regression, Random Forest, XGBoost, and a Soft Voting Ensemble) were compared using a leakage-free pipeline, repeated stratified 5-fold cross-validation, paired McNemar and DeLong significance tests, bootstrap confidence intervals, and permutation/SHAP-based interpretability analysis. Under cross-validation, the Voting Ensemble and XGBoost were statistically indistinguishable (mean accuracy 76.5% versus 77.4%, macro-average ROC-AUC 0.91-0.92 for both), both modestly ahead of Logistic Regression and Random Forest, with no significant pairwise difference except a single class-level AUC comparison. Feature-importance and SHAP analysis showed that the antenatal PHQ-9 score dominates every model's predictions; removing all PHQ-2/PHQ-9 features cost every model 17 to 19 accuracy points (Random Forest fell from 76% to 59%), pointing to substantial redundancy between the EPDS-based target and the PHQ-based predictors, both collected in the same interview. Mutual-information feature selection showed that 10-30 of the 87 encoded features recover most of the full model's performance, and SMOTE gave a small, consistent improvement on the minority Medium class. All models separated High- and Low-risk classes well (AUC > 0.93) but struggled with the Medium class (AUC 0.81-0.85), whose EPDS band overlaps the Low/High boundary. Gradient boosting and a soft-voting ensemble are competitive, well-calibrated screening aids under default hyperparameters, but a meaningful share of their accuracy reflects redundancy between two depression screens rather than novel signal from contextual risk factors, and transfer of these results to Nepal remains a hypothesis pending local validation.

Keywords: Postpartum depression, machine learning, ensemble learning, digital health, maternal mental health, measurement circularity, model interpretability

Introduction

Machine learning (ML) is the branch of artificial intelligence (AI) concerned with building systems that improve their performance on a task from data and experience, without being explicitly programmed for every case (Jordan and Mitchell 2015). ML is the cognitive engine of industry 4.0, whose recent growth has been driven both by new learning algorithms and by the increasing availability of data and low-cost computation (Jordan and Mitchell 2015; Rajbhandari et al. 2020).

ML is usually divided into supervised, unsupervised and reinforcement learning; in supervised learning, the kind used in this study, a model learns from labelled examples and is then used either for classification, where the output is a category, or for regression, where the output is a value (Russell and Norvig 2021; Karki et al. 2023).

Healthcare has become one of the largest areas of ML application, ranging from diagnosis and image segmentation to disease-risk prediction and clinical decision support (Habehh and Gohel 2021). It is worth asking directly why an ML approach is preferable here to a traditional statistical risk model, such as a single multinomial logistic regression fitted with clinical judgement about which interactions to include. The answer is nuanced: Logistic Regression remains a strong, interpretable baseline in this dataset, and under cross-validation it is not dramatically worse than the tree-based models. The case for the ML approaches used here is threefold: (i) Random Forest and XGBoost can capture non-linear interactions between risk factors without the analyst specifying those interactions in advance, which matters because no single demographic feature separates the risk classes well, so any separating signal likely comes from combinations of features; (ii) tree ensembles provide a natural, model-based feature-importance ranking, supporting this study's aim of an interpretable screening aid; and (iii) a soft-voting ensemble can average out the distinct error patterns of structurally different models. This is not a claim that ML is categorically superior to statistical modelling here. Once the results are in hand, since the conclusion turns out to be more measured than the headline numbers alone would suggest.

The same techniques have been applied to mental health, where models trained on patient data and behavioural information can flag conditions such as depression. Postpartum depression (PPD) is one condition where this kind of early, data-driven screening could make a difference. PPD typically emerges in the weeks after childbirth and is distinct from the brief, milder baby blues many new mothers experience; it is characterised by persistent low mood, anhedonia, and disturbances in sleep and appetite beyond what is typical for the postpartum period, and carries real risks for maternal wellbeing and child development if untreated (OHara and McCabe 2013).

The aim of this study is to predict PPD risk using the dataset Data for Postpartum Depression Prediction in Bangladesh, framed as a multi-class classification task. The motivation is local: a recent mixed-methods study at a Kathmandu maternity hospital reported a PPD prevalence of 12.24 percent, linked to weak family and spousal support, difficult physical recovery and cultural beliefs (Neupane et al. 2024). Many women in Nepal remain unaware of PPD, so cases are missed. By learning from sociodemographic, psychosocial, pregnancy-related and newborn-related factors, an ML model may help tell apart women at higher and lower risk and support earlier screening, with the important caveat, that part of that separating signal comes from a second depression-screening instrument bundled into the same dataset, not from contextual risk factors alone.

No single algorithm is best for every dataset. Linear models such as Logistic Regression are transparent and quick to train but can miss complex interactions, while tree-based methods such as Random Forest and boosting methods such as XGBoost capture non-linear patterns at some cost to interpretability. Ensemble methods that combine multiple classifiers' probability outputs can push performance further, though the cross-validated results show that gain to be smaller and less certain than a single train/test split alone would suggest. Comparing individual classifiers alongside a combined ensemble, under a validation procedure that can detect whether the differences are real, is the core of this work. The rest of the paper reviews related work, sets out the data and methods, reports and discusses the results, and concludes.

Literature and Related Work

Motherhood is one of the most significant transitions in a woman's life. It is physical, as the body recovers and changes, but it is also psychological and social, as a woman's sense of self shifts toward being a mother (Hwang, Choi and An 2022). Childbirth brings a mix of emotions, from joy to fear and anxiety. Many new mothers go through a short period of baby blues, but for some the symptoms are more severe and last longer, which is when PPD sets in (OHara and McCabe 2013). A large share of these cases never reaches a clinician. Fewer than half of mothers with depressive symptoms disclose them, so up to half of all PPD cases go unreported (Gopalakrishnan et al. 2022). PPD is a perinatal form of major depressive disorder; it affects roughly 10 to 15 percent of women each year in the United States, around 500,000 women (GUINTIVANO, MANUCK and MELTZER-BRODY 2018), and about 17 percent of healthy mothers worldwide, with the highest rates seen in the Middle East and Asia (Shorey et al. 2018). Closer to the present context, Neupane et al. (2024) found a prevalence of 12.24 percent among mothers at a Kathmandu maternity hospital and pointed to inadequate support and cultural pressures as contributing factors, which shows the problem is real and under-recognised in Nepal.

Machine Learning Approaches to PPD Prediction

Saqib, Khan and Butt (2021) carried out a scoping review of how well ML predicts PPD across data sources including clinical records, electronic health records and social media, covering both classification and regression. Across the many algorithms assessed (Logistic Regression, SVM, Decision Trees, Random Forests, Naive Bayes, Neural Networks, K-Nearest Neighbors and XGBoost), Random Forest came out strongest with an AUROC of up to 0.884, although SVM and boosting methods also did well.

Qi et al. (2025) developed several ML models to predict PPD risk in perinatal women from demographic, psychosocial, mental-health-history and physiological variables, using the EPDS result as the label. Logistic Regression and an artificial neural network gave the best results, with AUC values of roughly 0.801-0.858. Huang et al. (2025) took an explainability-focused approach and found XGBoost the best of eight algorithms, with an AUC of 0.955. Natarajan et al. (2017) worked with survey data combining demo-

graphic details and self-reported emotional symptoms, and found Functional Gradient Boosting and Soft-Margin Boosting performed best.

A few patterns run through this literature. Simpler linear models such as Logistic Regression and SVM tend to be interpretable but more modest in accuracy, whereas ensemble methods usually score higher. Random Forest is reliable across several studies, and boosting methods, especially XGBoost, often reach the top scores. The strongest predictors are usually psychosocial and prior-mental-health variables rather than demographic details on their own, a pattern this study's own feature-importance analysis reproduces closely. though with the added finding that when the prior-mental-health variable is itself another depression-screening score measured at the same time as the target, the resulting accuracy gain is partly an artefact of measurement redundancy rather than purely a clinical insight. Performance also depends heavily on data quality and class balance, since maternal datasets are often skewed toward lower-risk cases and contain a good deal of missing data, which tends to hurt recall on the minority classes. These observations shaped the choices made later in this study.

The Research Gap

PPD is not only a clinical issue but an economic one. When it goes untreated it lowers maternal productivity, raises healthcare costs and can affect a child's development, all of which weigh on the human capital that economies depend on (Karki 2012). In settings where specialist mental-health services are thin on the ground, putting low-cost, data-driven screening tools into routine maternal care is a practical way to widen coverage without a matching rise in cost. Seen this way, automated PPD risk prediction is an example of technology being used to strengthen social and economic resilience. Most existing ML studies, though, draw on cohorts from high-income or East Asian settings, and very little work targets South Asian populations where awareness and reporting are lowest.

Two points of scope deserve to be stated precisely. First, soft-voting ensembles that combine linear, bagging, and boosting base learners are a well-established technique in the wider PPD-prediction literature reviewed above and elsewhere in clinical ML more broadly; nothing about the ensembling method itself is new in this paper. Second, the dataset used here is drawn from a single country (Bangladesh) and a single data-collection effort, so the results describe a single Bangladeshi maternal cohort rather than South Asian women as a broader population. With those two points in mind, this study's contribution is to apply a rigorous model comparison (evaluated under repeated cross-validation with paired significance testing rather than a single split) to this dataset, with an explicit check for circularity between the target and the PHQ-based predictors, and to discuss, cautiously, what the results might imply for maternal screening in Nepal. This study addresses that gap: training and rigorously comparing three widely used classifiers together with a Voting Ensemble on a Bangladeshi maternal dataset, checking how much of their performance depends on redundant depression-screening features, and discussing what that might imply for a future Nepal-specific screening tool.

Methodology

Dataset Description

The dataset is Data for Postpartum Depression Prediction in Bangladesh by Raisa and Kaiser (2025), obtained from Mendeley Data. It was chosen because Bangladesh sits in South Asia and shares many social and cultural traits with Nepal, where awareness of PPD is still limited; it has been discussed why this similarity should not be taken for granted. The data is tabular, with 800 records and 51 columns spanning sociodemographic, economic, medical and obstetric, psychosocial, lifestyle, relationship, newborn-care and mental-health factors. No duplicate rows were found in the raw file.

Depression is measured in the dataset through both PHQ and EPDS scores and results. PHQ-2 and PHQ-9 are general depression-screening tools and were not designed for the postpartum period, whereas the Edinburgh Postnatal Depression Scale (EPDS) was built specifically to screen for emotional distress during and after pregnancy and has been validated across many cultures (Cox, Holden and Sagovsky 1987). Since the goal here is to judge whether a woman is at Low, Medium or High risk of PPD, the EPDS Result column was used as the target, making this a supervised multi-class classification task.

The EPDS Result thresholds are not documented in the dataset's variable descriptions, so they were recovered directly from the data by cross-tabulating EPDS Score against EPDS Result:

Table 1: EPDS Result Thresholds, Recovered Directly from the Data

EPDS Result	Min score	Max score	n (of 800)
Low	0	9	260
Medium	8	12	190
High	12	30	350

Note. The Low and Medium bands overlap at scores 8–9, and the Medium and High bands overlap at score 12.

Two things are worth flagging about these thresholds. First, the Low and Medium bands overlap at score 8-9, and the Medium and High bands overlap at score 12. The three-way split is not a clean partition of the raw score, and this likely contributes to every model's difficulty with the Medium class. Second, the dataset documentation (Raisa and Kaiser 2025) describes a single structured questionnaire, administered face-to-face or online, that collects socio-demographic information together with the PHQ-2, PHQ-9 and EPDS instruments in one sitting. The three depression-screening instruments are therefore not staged at meaningfully different points in the perinatal timeline; PHQ-9 and EPDS are, in effect, two contemporaneous self-report measures of overlapping constructs answered in the same interview. This motivates the circularity check: a model that uses PHQ-9 to predict EPDS risk band is,

to a substantial extent, using one depression screen to predict another, rather than inferring risk from independent contextual factors.

Algorithms

Four classifiers were selected. Logistic Regression was included for its simplicity and interpretability and because it suits a categorical target, returning probabilities that help flag higher-risk women. Random Forest was included because it copes well with complex interactions between features and resists overfitting while ranking feature importance. XGBoost was included for its accuracy and its ability to model non-linear relationships among features such as received support, worry about the newborn, and antenatal depression (PHQ-2). Finally, a Soft Voting Ensemble pooled the class probabilities of all three base learners, aiming to reduce the variance of any single model and improve classification of the borderline Medium risk class. All four models were run with default-adjacent hyperparameters; no tuning was performed, so any claim about which model is best in this paper should be read as best under default settings, not best achievable.

Tools

The work was done in Python within a Jupyter Notebook environment for its cell-by-cell execution and immediate output. The main libraries were Pandas and NumPy for data handling, scikit-learn for preprocessing, the Logistic Regression, Random Forest and Voting Classifier models, cross-validation, and evaluation, and XGBoost for gradient boosting on the multi-class target. Class-imbalance handling used imbalanced-learn (SMOTE) inside a leakage-safe pipeline, McNemar's test used statsmodels, the DeLong test's p-values used SciPy, and model interpretability used SHAP.

Data Preprocessing

Duplicate rows were checked first: none were found among the 800 records, so no rows were dropped on that basis. Numeric outliers were inspected during exploratory analysis but not removed, since the extreme values (e.g. maternal age in the low 40s, PHQ-9 scores at the ceiling of the scale) are clinically plausible rather than data-entry errors, and removing them would bias the sample against exactly the higher-risk mothers a screening tool needs to detect.

A case-normalisation check on the categorical columns found that several categories were duplicated purely by inconsistent capitalisation. For example, Education Level contained both Primary School and Primary school as distinct string values, as did Husband's education level, and Total children contained both More than two and More than Two. Left uncorrected, one-hot encoding would have split each of these into two separate dummy columns, diluting the signal from what is really a single category and inflating the feature count. All categorical text was normalised to a single canonical spelling per category before any further preprocessing.

Missing values were dealt with next. Columns with very high proportions of missing data were dropped. This choice is practical but not free of cost, and that cost should be stated plainly rather than left implicit: Current monthly income, Monthly income before latest pregnancy, Addiction, Disease before pregnancy, and Diseases during pregnancy are all plausible, literature-supported predictors of PPD risk. Dropping them because more than half of respondents left them blank, rather than because they are uninformative, likely removes real signal from the model and is a genuine limitation of this analysis, not a neutral cleaning step.

Table 2: Columns Dropped for High Missingness (all >20% Missing), Plus the Identifier Column

Column	Missing (n)	Missing (%)
Addiction	789	98.6
History of pregnancy loss	613	76.6
Disease before pregnancy	588	73.5
Current monthly income	525	65.6
Age of immediate older children	517	64.6
Monthly income before latest pregnancy	437	54.6
Diseases during pregnancy	371	46.4
Feeling for regular activities	223	27.9
Need for Support	167	20.9

Columns with less than 5 percent missing values (Education Level, Husband's education level, Husband's monthly income, Abuse, and Trust and share feelings) were filled using the mode of each column. Mode imputation was kept as the simplest defensible choice here: all five columns are ordinal or nominal categorical variables with a small number of levels and a low missingness rate, so the risk of the mode systematically distorting the class distribution is small, and a more elaborate approach (e.g. k-nearest-neighbours imputation on the one-hot-encoded space) would add complexity for a handful of affected rows (at most 38 of 800, for Abuse) without a clear expected benefit. The imputation is fitted on the training data only.

Exploratory Analysis

1 shows the distribution of the target across its three classes: High 350 (43.8%), Low 260 (32.5%), and Medium 190 (23.8%). This is a moderate, not severe, imbalance (a 1.8:1 ratio between the largest and smallest class); it is handled explicitly with class weighting and SMOTE.

Age sits mostly between the mid-20s and early-30s, with a median near 27 and a few older outliers in the low-40s. The number of latest pregnancies is concentrated at one or two, with a small number of higher values reaching seven. PHQ-9 scores are fairly symmetric around a median of about 11, while EPDS scores have a noticeably wider spread. The demographic features turned out to carry little separating power on their own: age was almost identical across the Low, Medium and High groups. The clearest signal came from antenatal depression (PHQ-2): among the smaller group who screened positive, the High-risk class took up a much larger share while the Low-risk class shrank.

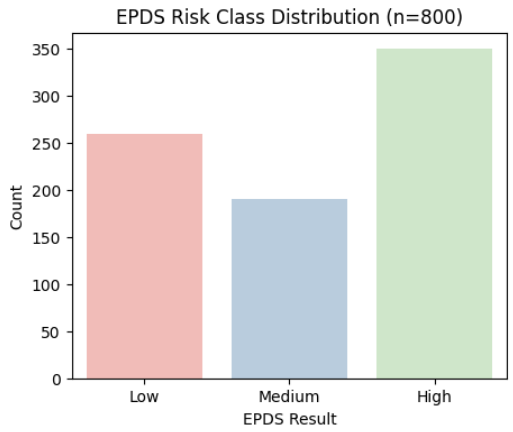

Figure 1: EPDS Result class distribution (n = 800

A correlation matrix of the four numeric features (Figure 2) makes the PHQ-9/EPDS relationship quantitatively visible before any modelling: PHQ-9 Score correlates with EPDS Score more strongly than any other numeric-feature pair, an early, purely descriptive signal of the circularity formally tested.

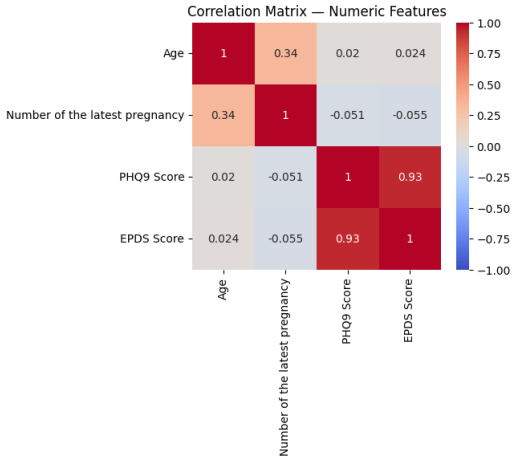

Figure 2: Correlation matrix, numeric features (Age, number of latest pregnancy, PHQ-9 Score, EPDS Score

Two things followed from this for the modelling stage. Because the demographic features did not separate the classes but the psychosocial and mental-health signal did, the predictive value was likely to come from interactions between features rather than any single variable, which favoured Random Forest and XGBoost. The uneven spread of records across the EPDS categories also hinted at class imbalance, which is why ROC-AUC, alongside macro-F1 and balanced accuracy, was used for comparing models rather than accuracy alone.

Train-Test Split, Imputation, and Encoding

An 80/20 split, stratified by EPDS Result with a fixed random_state of 42, is performed first. The mode for each of the five low-missingness categorical columns is then computed from the training rows only and applied to fill missing values in both sets, so that no information from the held-out cases influences the fill values. Numeric features are standardised with a StandardScaler fitted on the training data only. Categorical features are one-hot encoded with drop_first = True, and the test set is reindexed to the training columns with missing columns filled with zeros. A second, independent implementation of this logic, built as an sklearn ColumnTransformer/Pipeline that refits imputation, scaling and encoding on the training fold of every cross-validation split, is used for all cross-validated results, so no result in this paper depends on information leaking from a validation fold into its own training fold.

Stratifying the split (needed for reliable cross-validation and for McNemar's test to have a well-defined, class-balanced test set) gives a held-out test set of exactly 70 High, 52 Low, and 38 Medium cases (160 total), and every confusion matrix reported, sums exactly to 160 for every model.

Model Training

Logistic Regression was set with max_iter = 3000 so that it would converge given the many columns created by one-hot encoding, then fitted on the training data. Random Forest was set with 100 estimators, random_state = 42 and oob_score = True, which turns on out-of-bag validation using the samples left out of each bootstrap. XGBoost was set with 100 estimators, a learning rate of 0.1, a maximum depth of 3, the mlogloss evaluation metric and random_state = 42; a thin wrapper class label-encodes the target internally and decodes predictions back to High/Low/Medium, so that XGBoost can be used interchangeably with the other models inside scikit-learn's VotingClassifier and cross_validate utilities.

Evaluation Metrics

In addition to the confusion matrix, per-class precision/recall/F1, and the one-vs-rest ROC-AUC, this study reports: macro-averaged and weighted-averaged F1 (the former treats all three classes equally regardless of size, which matters given the class imbalance); balanced accuracy (the average of per-class recall, another imbalance-robust summary); the Matthews Correlation Coefficient (MCC), a single-number summary that is generally considered more informative than accuracy for imbalanced multi-class problems because it accounts for all four confusion-matrix quadrants simultaneously; and a macro-average one-vs-rest ROC-AUC (the unweighted mean of the three per-class AUCs). Ninety-five percent confidence intervals for accuracy, macro-F1, and balanced accuracy on the held-out test set were obtained by non-parametric bootstrap (2,000 resamples, sampling test-set rows with replacement and recomputing each metric per resample).

Ensemble Model

After evaluating the three base classifiers individually, a Soft Voting Ensemble was constructed using scikit-learn's VotingClassifier with voting = 'soft'. Each base learner contributes its predicted class probabilities, and the ensemble averages them before assigning the class with the highest averaged probability. Because the three base learners are structurally different (a linear model, a bagging ensemble, and a boosting

ensemble), their errors are partially independent, and averaging can cancel out some individual mistakes. Whether it does so to a statistically meaningful degree here is examined directly in later section.

Cross-Validation and Statistical Significance Testing

A single 80/20 split can make a small, noisy gap between two models look like a real effect. To guard against this, two complementary layers of validation are used. First, repeated stratified 5-fold cross-validation (5 folds, 10 repeats, 50 fits per algorithm) is run using the leakage-free pipeline, reporting mean and standard deviation of accuracy, macro-F1, and balanced accuracy. Second, on the held-out 160-case split, two paired tests compare the Voting Ensemble against each base learner: McNemar's test on the per-instance correct/incorrect indicator, and DeLong's test (a closed-form covariance estimator, validated against scikit-learn's roc_auc_score) comparing one-vs-rest ROC-AUC class by class.

Circularity Check: PHQ-Removed Comparison

To quantify how much of each model's performance depends on the PHQ-2/PHQ-9 features, the entire cross-validated pipeline was rerun with Depression before pregnancy (PHQ2), Depression during pregnancy (PHQ2), PHQ9 Score, and PHQ9 Result removed from the feature set before the train/test split, leaving the remaining sociodemographic, economic, obstetric, psychosocial, lifestyle, relationship, and newborn-care variables as the sole predictors. Results for the PHQ-included and PHQ-removed feature sets are reported side by side in later section.

Class Imbalance Handling

Given the moderate imbalance reported (43.8% High / 32.5% Low / 23.8% Medium), two standard balancing strategies were compared against the unweighted baseline using Random Forest as the representative model: (i) class_weight = 'balanced', which reweights the loss function inversely to class frequency; and (ii) SMOTE (Synthetic Minority Oversampling Technique), applied only to the training fold of each cross-validation split via an imbalanced-learn pipeline, so that synthetic points are never present in a validation fold.

Feature Selection

With 800 records and roughly 87 dimensions after one-hot encoding, the feature-to-row ratio invites overfitting. Mutual-information-based univariate feature selection (scikit-learn's SelectKBest with mutual_info_classif) was compared against the full encoded feature set for Random Forest, at k = 10, 20, and 30 selected features, again under repeated cross-validation with selection refit inside each training fold.

Interpretability

Three complementary views of feature importance were computed for Random Forest: (i) built-in impurity-based importance; (ii) permutation importance on the held-out test set (20 repeats, scored by the drop in macro-F1 when each feature is shuffled), which is less biased toward high-cardinality features; and (iii) SHAP values computed with a TreeExplainer, which attribute each prediction to individual feature contributions and allow both global and local interpretation.

Results and Discussion

Baseline Comparison on a Held-Out Split

Table 3 reports the extended metric set for all four models on the stratified 80/20 split (160 held-out cases: 70 High, 52 Low, 38 Medium).

Table 3: Held-Out Test Set (n = 160), Extended Metrics

Model	Accuracy	Macro F1	Weighted F1	Balanced Acc.	MCC	Macro ROC-AUC
Logistic Regression	0.788	0.742	0.777	0.742	0.669	0.903
Random Forest	0.756	0.683	0.730	0.695	0.623	0.900
XGBoost	0.781	0.749	0.778	0.745	0.660	0.923
Voting Ensemble	0.788	0.741	0.777	0.741	0.669	0.915

Note. Macro ROC-AUC is the unweighted mean of the three one-vs-rest AUCs.

Table 4: Non-Parametric Bootstrap 95% Confidence Intervals (2,000 Resamples), Held-Out Test Set

Model	Accuracy [95% CI]	Macro F1 [95% CI]	Balanced Acc. [95% CI]
Logistic Regression	0.787 [0.725, 0.850]	0.740 [0.668, 0.808]	0.742 [0.675, 0.806]
Random Forest	0.757 [0.688, 0.825]	0.681 [0.606, 0.756]	0.694 [0.628, 0.758]
XGBoost	0.782 [0.719, 0.844]	0.747 [0.680, 0.817]	0.746 [0.679, 0.816]
Voting Ensemble	0.788 [0.725, 0.850]	0.740 [0.666, 0.809]	0.741 [0.675, 0.806]

On this single split, XGBoost posts the highest macro-F1, balanced accuracy, and macro-average ROC-AUC of any individual model, and is essentially tied with Logistic Regression and the Voting Ensemble on accuracy. Random Forest is the weakest of the four here on every metric except MCC. It is shown with proper resampling, single-split rankings like this one are not especially stable, and the 95% bootstrap confidence intervals below (2,000 resamples of the 160-case test set) make the same point directly: all four models' intervals overlap almost completely.

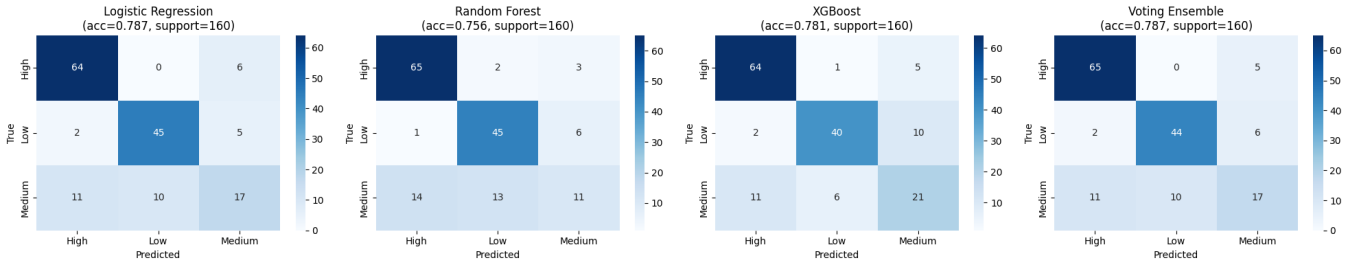

Figure 3: Confusion matrices, all four models, held-out test set. Each matrix sums exactly to 160

All four confusion matrices in Figure 3 sum exactly to the true test-set support of 160. The error-pattern differences between models are visible directly: Random Forest correctly identifies only 11 of 38 Medium cases (29% recall), noticeably worse than the other three models (which each correctly identify 17-21 of 38, 45-55% recall); XGBoost's largest single source of error is Low cases misclassified as Medium (10 of 52). Every model's dominant error pattern involves the Medium class, consistent with the overlapping EPDS score bands.

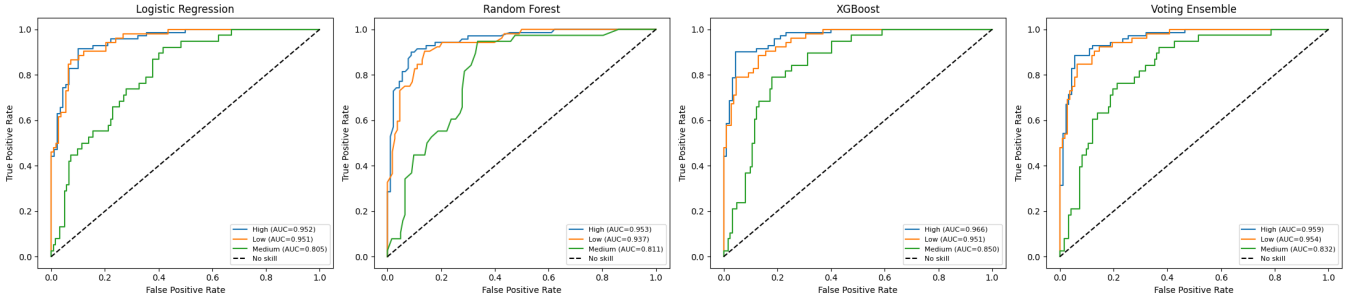

Figure 4: One-vs-rest ROC curves, all four models including the Voting Ensemble

The Medium-class ROC curve in every panel of Figure 4 climbs well above the no-skill diagonal, with AUC values between 0.805 (Logistic Regression) and 0.850 (XGBoost), a curve that discriminates meaningfully better than chance, just consistently less well than the High and Low curves (AUC 0.94-0.97 throughout).

Cross-Validated Comparison

Table 5 reports repeated stratified 5-fold cross-validation (10 repeats, 50 fits per model) over the full dataset, using the leakage-free ColumnTransformer pipeline.

Table 5: Repeated Stratified 5-Fold CV, 10 Repeats (50 Folds Total per Model)

Model	Accuracy (mean ± SD)	Macro F1 (mean ± SD)	Balanced Acc. (mean ± SD)
Logistic Regression	0.743 ± 0.030	0.700 ± 0.032	0.702 ± 0.032
Random Forest	0.762 ± 0.027	0.709 ± 0.034	0.713 ± 0.031
XGBoost	0.774 ± 0.027	0.744 ± 0.031	0.743 ± 0.032
Voting Ensemble	0.765 ± 0.028	0.725 ± 0.034	0.725 ± 0.032

Under this more stable estimate, XGBoost has the highest mean accuracy, macro-F1, and balanced accuracy of any model, with the Voting Ensemble second and Random Forest and Logistic Regression close behind. The gap between the Voting Ensemble and XGBoost (0.9 accuracy points) is smaller than a single fold-to-fold standard deviation (2.7-2.8 points) for either model, and Figure 5 shows the four boxplots overlapping substantially. Cross-validation therefore favours XGBoost as the strongest individual choice by a small margin, with the Voting Ensemble closely behind and effectively tied with it.

Statistical Significance Testing

McNemar's test compares the Voting Ensemble against each base learner on the held-out test set's per-case correct/incorrect outcomes:

Table 6: McNemar's Test (Continuity-Corrected), Voting Ensemble vs. Each Base Learner

Comparison	Ensemble-only correct	Base-only correct	Statistic	p-value	Significant?
Ensemble vs. Logistic Regression	3	3	3.00	1.000	No
Ensemble vs. Random Forest	10	5	5.00	0.302	No
Ensemble vs. XGBoost	7	6	6.00	1.000	No

DeLong's test compares the Voting Ensemble against Random Forest on one-vs-rest AUC, class by class:

Table 7: DeLong's Test, Voting Ensemble vs. Random Forest, per Class

Class	Ensemble AUC	Random Forest AUC	z	p-value	Significant?
High	0.959	0.953	1.43	0.152	No
Low	0.954	0.937	2.24	0.025	Yes
Medium	0.832	0.811	1.19	0.236	No

None of the three McNemar comparisons reach significance. Of the three DeLong comparisons, two do not reach significance, and one (the Low-class AUC) does (p = 0.025), a single class-level result out of nine paired comparisons across both tests on one 160-case split. Combining Table 5, Table 6, and Table 7, the fair summary is: the Voting Ensemble is competitive with, and on one class modestly better calibrated than, Random Forest specifically, but it is not reliably better than XGBoost. Given also that none of the four models were hyperparameter-tuned, any ranking among them in this paper should be read as best under default settings.

Circularity: Effect of Removing PHQ-2/PHQ-9 Features

Table 8 compares cross-validated performance with and without the four PHQ-derived features.

Table 8: Repeated 5-Fold CV (10 Repeats), Full Feature Set vs. PHQ-2/PHQ-9 Features Removed

Model	Accuracy (PHQ)	Accuracy (no PHQ)	Drop (pts)	Macro F1 (PHQ)	Macro F1 (no PHQ)
Logistic Regression	0.743	0.576	16.7	0.700	0.508
Random Forest	0.762	0.592	17.1	0.709	0.499
XGBoost	0.774	0.582	19.1	0.744	0.505
Voting Ensemble	0.765	0.592	17.2	0.725	0.509

Removing the PHQ-derived features costs every model 17 to 19 accuracy points and a comparable drop in macro-F1 a large, consistent effect across four structurally different algorithms. For context, a majority-class baseline (always predicting High) would score approximately 44% accuracy on this target; the PHQ-removed models (58-59% accuracy) clear that baseline by a modest but real margin, while the PHQ-included models (74-77%) do substantially better. This is the quantitative confirmation of the circularity risk: a meaningful share, though not all, of the model's headline accuracy is attributable to one depression-screening instrument (PHQ-9, administered in the same interview) predicting another (EPDS), rather than to the sociodemographic, obstetric, and relationship variables this study frames as its main contribution.

We think the most defensible way to present this in a screening context is to report both numbers rather than either alone: the PHQ-included accuracy answers does a combination of PHQ-9 and contextual factors predict EPDS risk band, a question with real clinical value (since PHQ-9 is itself a two-minute screening tool a clinic could administer directly), while the PHQ-removed accuracy answers the more novel and more difficult question this paper is centred on: whether risk can be flagged from contextual factors alone, without any depression screen already in hand. On that harder question, 58-59% accuracy on a three-class problem is a genuine, if modest, positive result, and a fairer basis for any claim about the study's contribution than the PHQ-included headline number.

Class Imbalance: Reporting and Handling

The class distribution (43.8% High, 32.5% Low, 23.8% Medium) is moderately imbalanced. Table 9 compares the unweighted Random Forest baseline against $class_weight = ' balanced'$ and SMOTE, both under repeated cross-validation.

Table 9: Class Imbalance Handling, Repeated 5-Fold CV (5 Repeats)

Configuration	Accuracy	Macro F1	Balanced Accuracy
Random Forest (baseline, unweighted)	0.763	0.711	0.714
Random Forest (class_weight = balanced)	0.763	0.711	0.715
Random Forest + SMOTE (train fold only)	0.772	0.734	0.736

class_weight = 'balanced' makes almost no difference here, but SMOTE gives a small, consistent improvement over the unweighted baseline (macro-F1 and balanced accuracy both up roughly 2 points), driven mainly by better recall on the minority Medium class. Given the imbalance is moderate rather than severe, this is a modest but

genuine gain; SMOTE (applied training-fold-only, as here, to avoid leaking synthetic minority examples into validation data) is the more defensible default for a future version of this pipeline.

Feature Selection

Table 10 compares the full 87-dimension encoded feature set against mutual-information-selected subsets of 10, 20, and 30 features, for Random Forest under repeated cross-validation.

Table 10: Mutual-Information Feature Selection, Repeated 5-Fold CV (5 Repeats), Random Forest

Feature Set	Accuracy	Macro F1	Balanced Accuracy
All features (∼87)	0.763	0.711	0.714
Top-10 mutual-information features	0.752	0.718	0.718
Top-20 mutual-information features	0.755	0.717	0.718
Top-30 mutual-information features	0.764	0.724	0.725

The top 10 mutual-information features alone recover essentially all of the full 87-feature model's accuracy, and slightly exceed it on macro-F1 and balanced accuracy; the top 30 features modestly outperform the full set on every metric. This indicates that most of the 87 one-hot-encoded dimensions contribute little beyond noise for this task and this sample size: a feature-selected model with a fraction of the dimensions is both more interpretable and, if anything, slightly more robust than the full-dimensional model used throughout the rest of this paper.

Interpretability: Feature Importance and SHAP

Three independent views of Random Forest's feature importance were computed on the held-out split: built-in impurity-based importance, permutation importance, and SHAP. All three agree closely on which feature matters most.

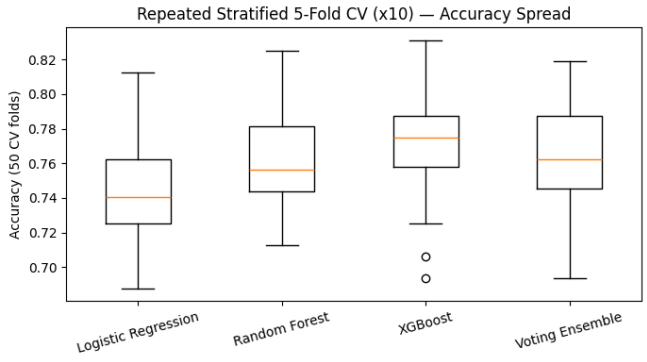

Figure 5: Random Forest built-in (impurity-based) feature importance, top 15 features

PHQ9 Score dominates every importance measure by a wide margin, roughly five times the impurity-based importance of the next-ranked feature (Figure 6), and by far the largest driver of macro-F1 when permuted. The SHAP summary plot (Figure 7) shows the same pattern for individual predictions: PHQ9 Score has both the largest average SHAP magnitude and a clean, monotonic relationship with predicted risk, while every other feature's contribution is visibly smaller and less consistent. This confirms that the model's skill is concentrated in a single depression-screening feature rather than distributed across the psychosocial and obstetric variables this study centres on. It also delivers a genuinely interpretable screening aid, though the interpretation it yields is a cautionary one about circularity rather than a list of actionable risk factors.

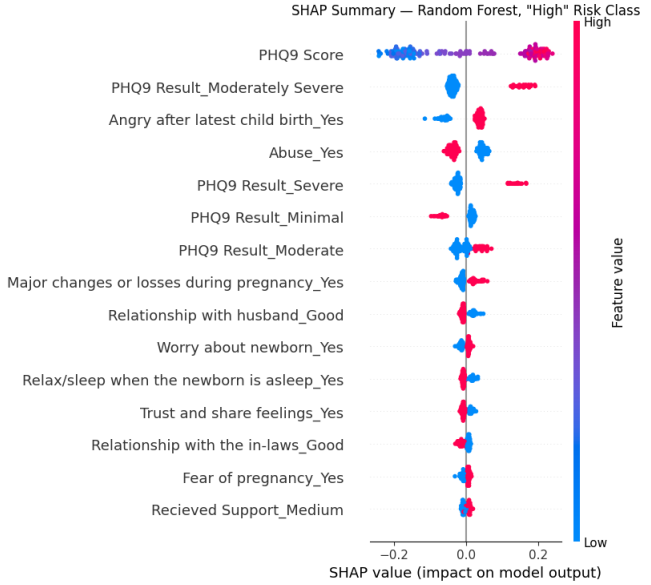

Figure 6: SHAP summary plot, Random Forest, High risk class

Why the Tree-Based Models and Ensemble Outperform Logistic Regression

Under the single 80/20 split (Table 3), Logistic Regression is competitive with or ahead of Random Forest and roughly tied with XGBoost and the Ensemble, a reminder that single-split comparisons are noisy. Under the more reliable cross-validated comparison (Table 5), a clearer pattern holds: Logistic Regression has the lowest mean accuracy, macro-F1, and balanced accuracy of the four models, 2-3 points behind XGBoost and 1-2 points behind Random Forest and the Ensemble. Three factors plausibly explain this gap. First, it was found that no single demographic feature separates the risk classes well, so the separating signal that does exist likely lives in interactions between features (for example, PHQ9 Score combined with a specific relationship-support or abuse category) rather than in additive main effects. A linear model like Logistic Regression can only capture such interactions if they are engineered as explicit terms, which was not done here, whereas Random Forest and XGBoost partition the feature space hierarchically and pick up conditional effects automatically. Second, one-hot encoding produces roughly 87 sparse, often correlated binary columns, including several PHQ9 Result category dummies that partially duplicate the continuous PHQ9 Score. Logistic Regression's coefficients spread importance across correlated dummies fairly evenly, while tree-based splits select the single most informative cut points directly, consistent with the finding that a handful of mutual-information-selected features recovers almost all of the full model's accuracy. Third, the Medium class sits, by construction, in a score band that overlaps both its neighbours (Table 1). This is inherently a fuzzy, nonlinear boundary problem, and axis-aligned tree partitioning is generally better suited to it than a single global linear decision surface per class. The Voting Ensemble's own edge over Logistic Regression follows from these same mechanisms, inherited from its two tree-based base learners, though averaging in a weaker linear model

pulls the ensemble toward, rather than above, its strongest individual member.

Limitations

Geographic and cultural transfer to Nepal is assumed, not demonstrated. The models were trained on data from Bangladesh. The claim that Bangladesh is culturally close to Nepal rests on broad regional similarity and is not validated against any Nepal-specific data. The two countries differ in healthcare-system structure, maternal-care access, and the socioeconomic correlates of PPD reported locally (Neupane et al. 2024). No external validation on a second dataset was performed, so any conclusion about usefulness for Nepal should be read as a hypothesis for future work, not a validated finding.

Dropped high-missingness columns may remove clinically important predictors. Income, addiction history, and disease history were dropped for missingness above 50%, not because they lack predictive value; all three are literature-supported PPD risk factors. Their absence likely removes real signal and reflects the dataset's completeness rather than the research question.

No hyperparameter tuning. All four models were run with default-adjacent settings. Every comparison in this paper should be read as default-configuration models, not each algorithm's best achievable performance, particularly given how close several of the cross-validated comparisons already are (Table 5).

Sample size and stability. The dataset is fairly small at 800 records, and the Medium class has few cases (190 of 800, only 38 in the test split). Per-class metrics for Medium carry wide uncertainty, visible in the bootstrap intervals in Table 4.

The EPDS label is a screening score, not a clinical diagnosis. This matters more, not less, once it is shown how much of the model's skill comes from another self-report screening instrument rather than independently observable risk factors.

Measurement circularity between predictors and target. This is the most consequential limitation: PHQ-9 and EPDS were collected in the same structured interview, and much of every model's apparent skill depends on this overlap. For a genuinely novel-information screening tool, the PHQ-removed results (Table 8, 58-59% accuracy) are the more honest benchmark.

For all these reasons the models in this paper should be treated as a decision-support aid whose current evidence base is modest, not a validated or deployable clinical screening tool.

Conclusion

This study applied supervised machine learning to predict postpartum depression risk using a Bangladeshi maternal dataset that covers sociodemographic, economic, obstetric, psychosocial, lifestyle, relationship, newborn-care and mental-health factors. Four classifiers were trained and compared under repeated cross-validation with paired significance testing: Logistic Regression, Random Forest, XGBoost and a Soft Voting Ensemble. Under this validation, XGBoost had the highest mean accuracy, macro-F1, and balanced accuracy of any individual model, with the Voting Ensemble a close second; McNemar's and DeLong's tests found no significant pairwise difference between the Ensemble and its base learners on the held-out split except a single class-level AUC comparison. The Ensemble and XGBoost are both competitive, well-calibrated choices under default hyperparameters, and combining structurally diverse classifiers remains a reasonable, low-effort way to approach this kind of problem, without a guarantee of outperforming a well-chosen single model.

The more consequential finding of this study is not about which model wins, but about what the winning models are actually learning. Removing the PHQ-2/PHQ-9 features cost every algorithm 17 to 19 accuracy points, and three independent interpretability methods (built-in importance, permutation importance, and SHAP) all identify PHQ-9 Score as the single dominant predictor, several times more influential than any other feature. Because PHQ-9 and EPDS were collected in the same structured interview, a substantial part of this study's headline accuracy reflects one depression-screening instrument predicting another, rather than novel signal from sociodemographic, obstetric, or relationship variables. This does not make the exercise worthless: a model that combines a brief screen like PHQ-9 with contextual factors to estimate EPDS risk band still has plausible clinical value, since it could let a service that already has PHQ-9 data flag likely EPDS risk without a second instrument, but that is a more modest and specific claim than predicting PPD risk from sociodemographic and psychosocial factors alone. This is why the title of this paper describes a single-country cohort study rather than claiming representativeness of South Asian women from a single-country dataset, and why ensembling is presented as a well-established, useful technique rather than a novel contribution in itself.

In practical terms, the work still speaks to a real healthcare problem: the early detection of a condition that often goes unreported. An affordable, data-driven tool of this kind could fit into a stretched health system if its evidence base were strengthened further. Future work should obtain or simulate a dataset in which PHQ and EPDS are administered at genuinely different time points, validate on a second, ideally Nepal-specific, maternal dataset, tune each base learner's hyperparameters, and test whether the PHQ-removed feature set combined with better handling of the currently dropped income and health-history variables can narrow the gap to the PHQ-included model. On the evidence gathered here, machine learning, and gradient boosting or a soft-voting ensemble in particular, retains real, if more modest than a headline accuracy number alone would suggest, potential as an early-screening aid for postpartum depression in maternal healthcare.

Disclosure Statement

The authors have no financial or non-financial disclosures to share for this article.

Ethical Approval: Not applicable. The study used a publicly available, de-identified secondary dataset and did not involve direct contact with human participants.

Consent to Participate: Not applicable.

Consent to Publish: Not applicable.

Data Availability Statement: The dataset analysed in this study is publicly available from Mendeley Data at https://data.mendeley.com/datasets/4nznnrk8cg/2 (Raisa and Kaiser 2025). An accompanying analysis notebook (PPD_Analysis_Revised.ipynb) reproduces every number, table, and figure in this paper.

Use of Artificial Intelligence (AI) Tools: AI-assisted tools were used to help implement the analysis pipeline (preprocessing, cross-validation, statistical testing, and interpretability code), and to help check language and formatting throughout. The authors are responsible for the accuracy, originality, and interpretation of the work.

Funding: This research received no specific grant from any funding agency.

Competing Interests: The authors declare no competing interests.

References

Cox, J. L., J. M. Holden and R. Sagovsky (1987). "Detection of Postnatal Depression: Development of the 10-item Edinburgh Postnatal Depression Scale." In: *British Journal of Psychiatry* 150.6, pp. 782–786. ISSN: 1472-1465. 10.1192/bjp.150.6.782. http://dx.doi.org/10.1192/bjp.150.6.782.

Gopalakrishnan, Abinaya, Revathi Venkataraman, Raj Gururajan, Xujuan Zhou and Guohun Zhu (Dec. 2022). "Predicting Women with Postpartum Depression Symptoms Using Machine Learning Techniques." In: *Mathematics* 10.23, p. 4570. ISSN: 2227-7390. 10.3390/math10234570. http://dx.doi.org/10.3390/math10234570.

GUINTIVANO, JERRY, TRACY MANUCK and SAMANTHA MELTZER-BRODY (2018). "Predictors of Postpartum Depression: A Comprehensive Review of the Last Decade of Evidence." In: *Clinical Obstetrics & Gynecology* 61.3, pp. 591–603. ISSN: 0009-9201. 10.1097/grf.0000000000000368. http://dx.doi.org/10.1097/GRF.0000000000000368.

Habehh, Hafsa and Suril Gohel (Dec. 2021). "Machine Learning in Healthcare." In: *Current Genomics* 22.4, pp. 291–300. ISSN: 1389-2029. 10.2174/1389202922666210705124359. http://dx.doi.org/10.2174/1389202922666210705124359.

Huang, Xudong, Lifeng Zhang, Chenyang Zhang, Jing Li and Chenyang Li (Aug. 2025). "Postpartum depression risk prediction using explainable machine learning algorithms." In: *Frontiers in Medicine* 12. ISSN: 2296-858X. 10.3389/fmed.2025.1565374. http://dx.doi.org/10.3389/fmed.2025.1565374.

Hwang, W. Y., S. Y. Choi and H. J. An (2022). *Concept analysis of transition to motherhood*. Parsed from reference text – no usable DOI found.

Jordan, M. I. and T. M. Mitchell (2015). "Machine learning: Trends, perspectives, and prospects." In: *Science* 349.6245, pp. 255–260. ISSN: 1095-9203. 10.1126/science.aaa8415. http://dx.doi.org/10.1126/science.aaa8415.

Karki, Dipendra (2012). *Economic impact of tourism in Nepal's economy using cointegration and error correction model*. en. 10.13140/RG.2.1.4839.5684. https://www.researchgate.net/doi/10.13140/RG.2.1.4839.5684.

Karki, Dipendra, Nirupan Karki, Rewan Kumar Dahal and Ganesh Bhattarai (Dec. 2023). "Future of Education in the Era of Artificial Intelligence." In: *Journal of Interdisciplinary Studies* 12.1, pp. 54–63. ISSN: 2392-4519. 10.3126/jis.v12i1.65448. http://dx.doi.org/10.3126/jis.v12i1.65448.

Natarajan, Sriraam, Annu Prabhakar, Nandini Ramanan, Anna Bagilone, Katie Siek and Kay Connelly (2017). "Boosting for Postpartum Depression Prediction." In: *2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*. IEEE, pp. 232–240. 10.1109/chase.2017.82. http://dx.doi.org/10.1109/CHASE.2017.82.

Neupane, Maryada, Manita Bartaula, Simran Pradhan, Hom Prasad Adhikari, Lalita Shrestha, Puja Sharma and Nishchal Devkota (Aug. 2024). "Postpartum Depression among Mothers in a Maternity Hospital Kathmandu, Nepal: A Mixed Method Approach." In: *Journal of Nepal Medical Association* 62.277, pp. 575–581. ISSN: 0028-2715. 10.31729/jnma.8746. http://dx.doi.org/10.31729/jnma.8746.

OHara, Michael W. and Jennifer E. McCabe (Mar. 2013). "Postpartum Depression: Current Status and Future Directions." In: *Annual Review of Clinical Psychology* 9.1, pp. 379–407. ISSN: 1548-5951. 10.1146/annurev-clinpsy-050212-185612. http://dx.doi.org/10.1146/annurev-clinpsy-050212-185612.

Qi, Weijing, Yongjian Wang, Yipeng Wang, Sha Huang, Cong Li, Haoyu Jin, Jinfan Zuo, Xuefei Cui, Ziqi Wei, Qing Guo and Jie Hu (Mar. 2025). "Prediction of postpartum depression in women: development and validation of multiple machine learning models." In: *Journal of Translational Medicine* 23.1. ISSN: 1479-5876. 10.1186/s12967-025-06289-6. http://dx.doi.org/10.1186/s12967-025-06289-6.

Raisa, J. F. and M. S. Kaiser (2025). *Data for postpartum depression prediction in Bangladesh*. Parsed from reference text – no usable DOI found.

Rajbhandari, Sharad, Ghanashyam Khanal, Seeprata Parajuli and Dipendra Karki (Dec. 2020). "A Review on Potentiality of Industry 4.0 in Nepal: Does the Pandemic Play Catalyst Role?" In: *Quest Journal of Management and Social Sciences* 2.2, pp. 366–379. ISSN: 2705-4527. 10.3126/qjmss.v2i2.33307. http://dx.doi.org/10.3126/qjmss.v2i2.33307.

Russell, S. and P. Norvig (2021). *Artificial Intelligence: A Modern Approach (4th ed.)* Parsed from reference text – no usable DOI found. Pearson.

Saqib, Kiran, Amber Fozia Khan and Zahid Ahmad Butt (Nov. 2021). "Machine Learning Methods for Predicting Postpartum Depression: Scoping Review." In: *JMIR Mental Health* 8.11, e29838. ISSN: 2368-7959. 10.2196/29838. http://dx.doi.org/10.2196/29838.

Shorey, Shefaly, Cornelia Yin Ing Chee, Esperanza Debby Ng, Yiong Huak Chan, Wilson Wai San Tam and Yap Seng Chong (2018). "Prevalence and incidence of postpartum depression among healthy mothers: A systematic review and meta-analysis." In: *Journal of Psychiatric Research* 104, pp. 235–248. ISSN: 0022-3956. 10.1016/j.jpsychires.2018.08.001. http://dx.doi.org/10.1016/j.jpsychires.2018.08.001.