Machine Learning–Based Prediction of High-Risk Pregnancies Using Maternal Health Indicators

– by Aasma Gupta (00601012024) & Akshata Jaiswal (01201012024)

1. Objective

The objective of this study is to develop a robust machine learning model capable of accurately predicting high-risk pregnancies based on maternal health indicators. The proposed system leverages key clinical parameters, including systolic and diastolic blood pressure, blood sugar levels, body temperature, heart rate, and maternal age, to classify pregnancy risk into predefined categories. By implementing data preprocessing, feature scaling, and supervised learning algorithms, the study aims to enhance the precision and reliability of risk assessment. The ultimate goal is to facilitate early detection of potential complications, thereby enabling timely medical intervention, reducing maternal and neonatal mortality, and supporting data-driven decision-making in obstetric healthcare practices.

2. Methodology

2.1 Model Architecture

- **Random Forest Classifier** – Selected as the primary model for its robustness, ability to handle non-linear relationships, and suitability for tabular medical data.
- Logistic Regression – Used initially for baseline interpretability.

2.2 Tools & Configuration

- **Programming Language**: Python
- **Libraries**: pandas, numpy, scikit-learn, matplotlib, seaborn
- **Explainability Tool**: Feature importance analysis from Random Forest for understanding variable contribution.

2.3 Data Cleaning & Integration

- **Dataset**: Maternal Health Risk Dataset containing Age, Systolic BP, Diastolic BP, Blood Sugar, Body Temperature, and Heart Rate.

2.4 Evaluation Metrics

- **Accuracy** – Overall prediction correctness.
- **Precision & Recall** – Class-specific control of false positives and false negatives, crucial for “High Risk” detection.
- **F1-Score** – Balanced metric accounting for both precision and recall.
- **Confusion Matrix** – Class-wise performance visualization for better interpretability.
- **Feature Importance** – Identification of the most influential maternal health indicators for prediction.

3. Dataset Used

Source: Maternal Health Risk Dataset (publicly available on Kaggle).
Size: 1,014 records.
Attributes:

- **Age** – Maternal age in years.
- **SystolicBP** – Systolic blood pressure (mmHg).
- **DiastolicBP** – Diastolic blood pressure (mmHg).
- **BS** – Blood sugar level (mmol/L).
- **BodyTemp** – Body temperature (°C).
- **HeartRate** – Heart rate (beats per minute).
- **RiskLevel** – Target variable with three classes: *Low Risk*, *Mid Risk*, and *High Risk*.
- **Class Distribution**: Balanced distribution maintained across *Low Risk*, *Mid Risk*, and *High Risk* categories through stratified sampling during train-test split.
- **Data Quality**: No significant missing values; minor inconsistencies were resolved during preprocessing.

4. Expected Outcomes

- **Accurate Risk Classification** – The system is expected to reliably categorize pregnancies into *Low Risk*, *Mid Risk*, and *High Risk*, even when applied to data from rural healthcare settings.
- **Early Detection in Low-Resource Areas** – Timely identification of high-risk cases where access to advanced diagnostic facilities is limited, enabling preventive action through community health workers.
- **Feature Insights for Local Healthcare** – Highlighting the most critical maternal health indicators for rural populations, such as elevated blood pressure or abnormal blood sugar, to guide targeted awareness campaigns.
- **Scalable and Low-Cost Deployment** – The model's lightweight design allows integration into mobile or tablet-based applications for use in primary health centers and field clinics.
- **Support for Government Programs** – Potential alignment with national maternal health initiatives to improve outcomes in underserved regions.