Project Discussion

Introduction

This project used the Maternal Health Risk Dataset from the UCI Machine Learning Repository. The data were collected from various hospitals and clinics in Bangladesh and comprise 1,013 observations with six predictor variables—Age, SystolicBP, DiastolicBP, BS (Blood Sugar), BodyTemp (Body Temperature), and HeartRate—along with the target variable RiskLevel, which represents the maternal mortality risk level.

Because the target variable is categorical, all models were implemented as classification models rather than regression models. Regression models predict continuous outcomes, while classification models predict discrete categories. Models such as Random Forest Classifier, XGBoost, LightGBM, and CatBoost were better suited for this task, as they can accurately model complex relationships between clinical features and categorical outcomes.

Results

The top-performing model was XGBoost. Among all models evaluated, it achieved the highest scores in accuracy, precision, macro-averaged F1, and ROC AUC, demonstrating strong classification performance across all risk categories. Although LightGBM showed a slightly better confusion matrix by one correct prediction, XGBoost consistently outperformed it in every other key metric. Its ROC AUC of 0.9379 further confirms its strong ability to distinguish between classes, as shown in the ROC curve.

Public Health Implications

These results demonstrate that models such as XGBoost and LightGBM can effectively predict maternal health risk using basic demographic and vital sign information. Machine learning models like these can help healthcare providers identify high-risk individuals before complications occur, enabling earlier intervention and potentially saving lives.

Importantly, the models performed well using easily obtained clinical features, such as blood pressure and heart rate—data typically collected during routine checkups. This finding highlights the potential of routine health screenings to inform maternal risk classification. Given that the dataset was derived from a low-income country, the results also suggest that predictive modeling could be a valuable tool in under-resourced and high-burden healthcare settings, supporting targeted interventions and resource allocation.

Caveats and Alternatives

The accuracy of any machine learning model depends on the quality and completeness of the data. Missing or biased data can lead to inaccurate predictions. The dataset used here contained no missing values, but potential biases were not assessed. This is an important consideration for future work.

Another limitation is the dataset’s limited feature set. Important factors such as socio-economic status, healthcare access, environmental exposures, and mental health were not included. These social determinants of health are known to strongly influence maternal outcomes and could enhance predictive accuracy if incorporated.

Potential improvements to this project include feature engineering or feature selection, which identify the most important predictors of the target variable. While this dataset contained only six features,future studies with richer datasets could benefit from these approaches. Additionally, ensemble methods that combine multiple models could further improve predictive performance by leveraging the strengths of different algorithms.