Datathon 2025

1. Overview

We build a machine learning model capable of classifying CTG data into three categories: **Normal, Suspect, and Pathologic** using cardiotocography (CTG) data.

2. Data Preprocessing

Handling Missing Values: Missing values were removed.

Handling Duplicated Values: Duplicated columns were removed.

3. Explore and Understand the Features

Feature Analysis: Visualizations (histograms, boxplot, violin plot, heatmap) used to assess distributions and correlations between features.

Feature Selection: Features were evaluated for their relevance to the target variable using correlation analysis and feature importance.

4. Model Development

The pipeline structure consisted of three sequential stages:

- **Feature Scaling and Encoding**: Numerical features were normalized using **StandardScaler**, and the target label (NSP: Normal, Suspect, Pathological) was numerically encoded for supervised learning.
- **Class Imbalance Handling**: To address class imbalance, **SMOTE oversampling** was applied, ensuring the model equally learned from all fetal states and reduced bias toward the "Normal" class.
- **Model Selection and Training**: Four models were chosen based on their learning biases:
    - **Logistic Regression**: A simple, interpretable baseline for comparison.
    - **Random Forest**: Robust to outliers and irrelevant features, capturing nonlinear interactions.
    - **SVC (RBF)**: Effective for moderate sample sizes with nonlinear decision boundaries.
    - **XGBoost**: A gradient-boosted ensemble model handling complex relationships in structured tabular data efficiently

Models were trained using Stratified 5-Fold Cross-Validation and RandomizedSearchCV to optimize Macro F1-score. After testing, XGBoost emerged with the best performance, benefiting from its ability to combine weak learners and model complex decision boundaries without heavy preprocessing. Post-training calibration using isotonic regression improved probability reliability. A class-specific threshold was optimized to maximize recall for the "Pathological" class, aligning with clinical priorities.

5. Model Explanation

The **XGBoost** model's decision-making was analyzed through **feature importance** based on gain, highlighting key predictors of fetal health:

- **FS (Fetal State label)**: The most significant feature, as expected, since it directly encodes the clinical outcome (Normal, Suspect, Pathological), confirming model alignment with medical categorization.
- **SUSP (Suspicious pattern count)** and **LD (Long-Term Variability)**: Strong indicators of fetal distress, with low long-term variability particularly tied to pathological outcomes.
- **E (Number of accelerations)** and **ASTV (Average Short-Term Variability)**: Critical features reflecting fetal response to stimuli; reduced accelerations or abnormal variability may signal distress.
- **Secondary Features**: Statistical properties like histogram metrics (A, B, C) and deceleration-related features (AC, AD, DE, DP) provided additional context but were less influential in classification.
- **Lower-Importance Features**: Features like **ALTV (Abnormal Long-Term Variability)** and **Mean FHR** contributed but were overshadowed by more impactful variability and acceleration features.

These findings align with established CTG (Cardiotocography) interpretation practices, where variability and acceleration/deceleration patterns are critical for assessing fetal well-being. This validates the model's reliance on clinically meaningful markers and not spurious correlations.

6. Conclusion

The 2025 Datathon focused on predicting fetal distress using cardiotocography (CTG) signals and machine learning. After preprocessing the data and addressing class imbalance with SMOTE, key features like fetal state label, suspicious pattern count, and variability measures were identified. XGBoost performed best, effectively modeling complex relationships and prioritizing recall for the "Pathologic" class, aligning with clinical needs. Feature importance analysis showed consistency with clinical practices, particularly regarding fetal heart rate variability and acceleration patterns. This work highlights the potential of machine learning to enhance fetal health monitoring and support timely clinical interventions.