Predicting Fetal Distress Using Cardiotocography Data

Objective: develop a machine learning model that assists clinicians in identifying fetal distress using real-world cardiotocography (CTG) recordings. The model classifies each CTG record into one of three risk categories—Normal, Suspect, or Pathologic—to support timely clinical decisions. The goal is to create a system that is not only accurate but also interpretable and applicable in real-world healthcare settings.

Data and Preprocessing Pipeline:

First, we obtain the data from the given link (Dua & Graff, 2019). Data cleaning involved checking for missing values and duplicate rows. Since the dataset is largely complete, there were no missing values, and a few duplicate rows were removed. Next, after splitting the data to test and train, feature scaling was performed to normalise feature distributions and prevent dominance of variables.

Model Design and Rationale:

We compared a few algorithm models to find the best-performing one: Gradient Boosting.

Logistic Regression was selected as a baseline due to its interpretability and clinical transparency. Decision Trees were included for feature importance visualisation, helping clinicians understand which CTG parameters influence outcomes most strongly. XGBoost was explored for potential performance gains through ensemble learning.

Model evaluation used recall and F1-score as key metrics to prioritise identifying pathologic cases—minimising false negatives, which are clinically critical. Cross-validation and confusion matrices were used to assess performance consistency.

Reference

Dua, D., & Graff, C. (2019). *Cardiotocography* [Data set]. UCI Machine Learning Repository. <https://archive.ics.uci.edu/ml/datasets/cardiotocography>