CTG Classification Model Development Report

This report summarizes the thought process and steps taken during the design, training, and optimization of machine learning models for the Datathon 2025 project. The objective was to classify fetal health conditions (Normal, Suspect, Pathologic) from Cardiotocography (CTG) data using various machine learning techniques.

**1. Data Understanding & Exploration**

The CTG dataset contained multiple features describing fetal heart rate patterns and uterine contractions. Initial exploratory analysis revealed that several features, such as ASTV and ALTV, were heavily skewed. To address this, a logarithmic transformation (log(x + 1)) was applied to normalize their distribution.

Correlation analysis showed that features like ‘Mean’, ‘Median’, ‘Mode’, and ‘Variance’ had moderate correlations with the fetal state (NSP), suggesting they are key indicators for classification. A heatmap of feature correlations was generated to visualize these relationships.

**2. Model Design and Training**

Multiple models were trained to determine the most effective architecture for this problem. These included Logistic Regression, Random Forest, XGBoost, LightGBM, CatBoost, and a Neural Network (MLP). To ensure fair evaluation, all models were trained using the same train-test split (80/20) and evaluated using Balanced Accuracy and Macro F1 Score metrics.

**3. Enhancements and Optimization**

The dataset exhibited class imbalance, with ‘Normal’ cases far outnumbering ‘Suspect’ and ‘Pathologic’ ones. To mitigate this, the Synthetic Minority Oversampling Technique (SMOTE) was applied to oversample minority classes. This significantly improved the model’s sensitivity to underrepresented classes, especially for Neural Networks.

Feature scaling was applied using StandardScaler and MinMaxScaler to stabilize gradient descent and improve model convergence. Tree-based models (Random Forest, XGBoost, CatBoost) achieved the highest accuracy, while Logistic Regression provided interpretability and Neural Networks captured complex relationships.

**4. Results and Insights**

(Fill in with updated results)

**5. Key Visualizations**

The following visualizations were used during analysis to better understand feature relationships and model behavior:

• Feature correlation heatmap (ASTV, ALTV, Mean, Median, Variance)

• Class distribution plot showing dataset imbalance

• Confusion matrix comparing predicted vs actual NSP classes

• Model comparison bar chart of Balanced Accuracy and F1 Scores  
  
(ADD THE CHARTS)