Fetal Health Classification from Cardiotocogram Data

Team name and SRN:-

Ishaan Sinha PES2UG23CS921

Ishika Raj PES2UG23CS923

Section J

Problem Statement

Classify fetal health status (normal, suspect, pathological) from cardiotocographic (CTG) measurements. This helps obstetricians assess fetal well-being during pregnancy and delivery, enabling timely interventions when necessary.

Dataset Description

This project utilizes the **Cardiotocography Data Set** from the UCI Machine Learning Repository. The dataset comprises 2,126 records of CTG examinations, each with multiple features used for prediction.

- **What the Dataset Contains**: The data consists of 21 numerical features extracted from CTG measurements.
- **Fields Used for Prediction**: The model uses all 21 measurement fields as input to predict the fetal health status. Key fields include baseline value (Fetal Heart Rate), accelerations, fetal_movement, uterine_contractions, and various histogram-derived statistics.
- **What the Model is Predicting**: The model predicts a single output variable, fetal_health, which is categorized into three classes:
  - **1 - Normal**
  - **2 - Suspect**
  - **3 - Pathological**

Project Methodology

The project was executed through distinct stages of data preparation and model development to ensure a robust and reliable outcome.

1 Data Preprocessing

Before model training, the dataset was carefully prepared using the following methods:

- **Outlier Removal**: To improve model performance by removing potentially erroneous data, we identified and removed extreme outliers using the **Z-score method**. Any data point where a feature had a Z-score greater than 3 was filtered out.
- **Data Splitting**: The cleaned dataset was split into two separate parts: an **80% training set**, used for teaching the models, and a **20% testing set**, reserved for evaluating their performance on unseen data.
- **Feature Scaling**: To prevent features with larger numerical ranges from disproportionately influencing the model, all features were scaled using the **StandardScaler**. This process standardizes each feature to have a mean of 0 and a standard deviation of 1.
- **Label Adjustment**: The target labels (1, 2, 3) were converted to a zero-indexed format (0, 1, 2) to ensure compatibility with the machine learning libraries.

2 Model Training

XGBoost (eXtreme Gradient Boosting): This is an advanced implementation of the gradient boosting algorithm. It builds decision trees sequentially, with each new tree specifically designed to correct the errors made by the previous one. To address the imbalanced nature of the dataset (where "Normal" cases far outnumber "Pathological" ones), **class weighting** was applied during training. This technique forces the model to place more importance on the rarer, critical cases.

Results and Evaluation

The performance of the XGBoost models was evaluated on the unseen test set using three key metrics.

- **Multi-class Accuracy**: This metric provides the overall percentage of correct predictions the model made across all three classes.
- **Per-class Precision, Recall, and F1-Score**: Presented in a detailed **Classification Report**, these metrics offer deeper insight into the model's performance on each specific class (Normal, Suspect, and Pathological). This is crucial for understanding if the model effectively identifies the rare but critical "Pathological" cases.
- **Confusion Matrix**: A visual table that provides a detailed breakdown of correct and incorrect predictions for each class. It clearly illustrates where the model is succeeding and where it might be making errors (e.g., confusing "Suspect" with "Normal").

Final Conclusion

XGBoost models demonstrated strong performance in classifying fetal health. After implementing class weight balancing for the XGBoost model to address data imbalance, it emerged as the top-performing model.

- The **XGBoost model achieved a final accuracy of approximately 95.6%**.

The high accuracy scores, combined with strong per-class metrics for the critical "Pathological" category, confirm that machine learning is a highly effective approach for this problem. The XGBoost model, in particular, stands out as a robust tool that could provide valuable, data-driven support to medical professionals in the crucial task of ensuring fetal well-being.