**DATATHON 2025**

**TM 210 - Academic Report**

**Duong Kien Khai, Stephen Manickam S/O Alakesan, Chong Jin Wei**

# 1. Exploratory analysis

By performing statistical tests and visualizations on the dataset, our team arrived at observations that aligned with known clinical interpretations, for example:

- High ASTV indicated reduced fetal responsiveness, hence a higher risk of distress.

- ASTV values were the lowest for Normal cases, higher for Suspect cases, and highest for Pathological cases.

# 2. Experimentation & Model choice

After testing with various models, we recorded that random forest and gradient boosting yield the best baseline results. We implemented these two methods with further optimizations and concluded that gradient boosting had the most prospect to enhance evaluation metrics scores and overall prediction results. Therefore, this technique is the focus of our team’s methodology.

We considered three gradient boosting models: XGBoost, CatBoost, and LightGBM. Since CatBoost “is designed to handle categorical features efficiently” and LightGBM “focuses on efficiency and scalability, particularly for large datasets and high-dimensional features” (A, 2024), while XGBoost is robust and efficiently versatile, our team decided to implement XGBoost as the main model.

# 3. Model pipeline

## 3.1. Data pre-processing

After importing the dataset, which is the ‘Data’ subsheet of ‘CTG.xls’, we performed an initial clean-up of the data in four sequential steps:

- Dropping the empty columns among the table,

- Dropping the columns with normalized data of a duplicated feature,

- Dropping the label-leakage columns, and

- Dropping row duplicates.

The cleaned dataset had a dimension of (2115, 25) — 2115 rows and 25 columns, of which 24 were feature columns and the remaining was the target column. Subsequently, we converted all data into a numerical format and relabel the classes as:

- ‘Normal’: 0 (78%)

- ‘Suspect’: 1 (14%)

- ‘Pathologic’: 2 (8%)

The fully pre-processed data contained no missing values.

## 3.2. Preparing training-testing data

Observing a significant imbalance in the target distribution, we implemented the conventional 80-20 split on the original dataset and stratified to ensure a similar proportion of all three classes on the training set and testing set. SMOTE was utilized to enhance the training set by increasing the presence of minority classes while minimizing the risk of oversampling. Finally, we achieved the training and testing sets with the following statistics:

*Training set:*

- ‘Normal’ / 0: 1318 instances

- ‘Suspect’ / 1: 1318 instances

- ‘Pathologic’ / 2: 1318 instances

*Testing set:*

- ‘Normal’ / 0: 329 instances

- ‘Suspect’ / 1: 59 instances

- ‘Pathologic’ / 2: 35 instances

## 3.3. Fitting the model

Fitting a baseline XGBoost model with n\_estimator=100, we achieved the following results:

- Balanced accuracy: 0.925

- Macro F1-score: 0.915

We used RandomizedSearchCV with 20 iterations to find the optimal hyperparameters for maximizing Balanced Accuracy and Macro F1-score, both of which gave the same set of optimal hyperparameters. We used these hyperparameters to construct and fit our optimal model, which is our main model in Datathon 2025, giving the following results:

- Balanced accuracy: 0.926

- Macro F1-score: 0.938

## 3.4. Explaining the model

We plotted each feature against its importance — how much the corresponding feature improves the XGBoost model’s performance when it is used in a split — and the SHapley Additive exPlanations (SHAP) values for each class, which measures how much each feature contributed (positively and negatively) to the corresponding class prediction. We were able to confirm that the model conformed to known clinical methodology of fetus health classification.

**Reference**

1. A, K. (2024, September 18). XGBoost vs. CatBoost vs. LightGBM: A Guide to Boosting Algorithms. *Medium*. https://kishanakbari.medium.com/xgboost-vs-catboost-vs-lightgbm-a-guide-to-boosting-algorithms-47d40d944dab