**Datathon 2025 – Lifeline**

**Team 76 | Repo:** [**https://github.com/vanessarifianto/datathon-2025-lifeline-team76**](https://github.com/vanessarifianto/datathon-2025-lifeline-team76%20) **| Academic Report**

**Problem Statement**

We frame CTG interpretation as a 3-class classification task: predict *Normal*, *Suspect,* or *Pathologic f*rom fetal heart rate and uterine contraction features. The aim is to support clinicians by detecting at-risk traces early while remaining interpretable (Ayres-de-Campos et al., 2000).

**Data Processing Pipeline**

We used the UCI Cardiotocography dataset (2,126 samples) labeled by obstetricians (UCI Machine Learning Repository, 2000). Twenty-one features were selected, including baseline FHR, accelerations, variability indices (ASTV, ALTV, mSTV, mLTV), and decelerations (DL, DS, DP). The pipeline included:

* Cleaning: dropped missing values and normalized headers.
* Transformation: mapped labels {1,2,3} → {0,1,2}; standardized features with StandardScaler.
* Preparation: addressed class imbalance using SMOTE oversampling inside the training pipeline to prevent leakage (Chawla et al., 2002).

**Model Design & Rationale**

We implemented a leak-free pipeline (*Scaler → SMOTE → Classifier*) evaluated with 5-fold stratified cross-validation. Logistic Regression served as a transparent baseline, while Random Forest was chosen for its ability to capture non-linear interactions and provide feature importances. A GridSearchCV tuned Random Forest hyperparameters (estimators, depth, splits), optimizing for macro-F1, a fairer metric for imbalanced classification.

**Rationale**

* Embedding preprocessing in the pipeline avoids data leakage.
* SMOTE ensures under-represented Pathologic cases are learned effectively.
* Macro-F1 and Balanced Accuracy give balanced evaluation across classes.
* Logistic Regression offers interpretability; Random Forest balances performance and explainability.

**References**

Ayres-de-Campos, D., Bernardes, J., Garrido, A., Marques-de-Sá, J., & Pereira-Leite, L. (2000). SisPorto 2.0: A program for automated analysis of cardiotocograms. *Journal of Maternal-Fetal Medicine*, 9(5), 311–318. [https://doi.org/10.1002/1520-6661(200009/10)9:5](https://doi.org/10.1002/1520-6661(200009/10)9:5%3c311::AID-MFM2%3e3.0.CO;2-9)

[<311::AID-MFM2>3.0.CO;2-9](https://doi.org/10.1002/1520-6661(200009/10)9:5%3c311::AID-MFM2%3e3.0.CO;2-9)

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357. <https://doi.org/10.1613/jair.953>

UCI Machine Learning Repository. (2000). *Cardiotocography dataset*. University of California, Irvine. Retrieved October 2025 from <https://archive.ics.uci.edu/dataset/193/cardiotocography>