Group: TM95 (Lee Hui Xin, Estelle & Alva Mridha & Adeline Nissi John)

Date: 02 October 2025

# **MLDA Datathon 2025 Report**

# 

# **Data Cleansing:**

In the original dataset, to tackle the issue of missing values, we replaced them with the median of the non-missing values and dropped the rows with missing labels to avoid errors. While this is more computationally efficient, it reduces variability in the dataset. In addition, for duplicates and redundant labels, we removed the redundant one-hot encodings. This prevents leakages as it prevents perfect multicollinearity, where one column can be exactly determined, resulting in inflated performance for models that assume feature independence.

**Exploratory Data Analysis:**

Through Exploratory Data Analysis, we can explore trends among Normal, Suspect and Pathologic cases. From the histogram of NSP against mSTV, it is observed that mSTV is lower in suspect/pathologic cases, which can be used for prediction. However, mLTV overlaps heavily with normal cases and is less discriminative. *(Appendix A, Fig 2)*  Boxplots of deceleration features revealed that DL exhibited the widest spread in pathological cases, reflecting more frequent and pronounced dips, whereas DS presented occasional outliers but no consistent class-specific trends, implying limited predictive strength. By contrast, DP was more common in pathological cases, making it a strong discriminator of fetal distress. *(Appendix A, Fig 3)* Statistical testing using the Kruskal–Wallis H test confirmed that features with low p-values, which are AC, ASTV/ALTV, and DP, differ significantly across classes, making them key predictors in distinguishing NSP.

**Feature engineering:**

Since there is a greater proportion of normal to suspect/pathologic cases in the dataset, models would be biased towards the majority classes. This leads to false negatives, which is especially worrying in this case. To combat this, we used class weights - assigning higher weights to underrepresented classes, thus creating a dictionary where each class label was mapped - and SMOTE oversampling, which artificially increases the number of minority class samples. For feature selection, we used the filter method, which evaluates each feature’s statistical association with its target variable, and selects the strongest. While efficient, this technique does not consider feature interactions nor model performance.

**Model:**

We chose Gradient Boosting as our model, as it can detect rare but critical cases and produce calibrated risk scores and supports class weighting, improving its performance on imbalanced datasets. This is crucial given we have more normal cases than suspect or pathological cases. Gradient Boosting is also able to capture subtle, non-linear feature interactions such as the outliers in features like deceleration patterns (DP), making it more effective compared to models like logistic regression. Given the high stakes of false negatives in the clinical context, this focus on rare cases is essential, even at the cost of more false positives. Gradient Boosting also directly optimizes for probability estimates, allowing it to support clinical decision-making processes by providing doctors with degrees of risk as opposed to binary outputs. Despite being more complex, it still offers insight into which features most influence prediction. This makes it more interpretable than neural networks, providing the degree of transparency clinicians need to understand and trust the model’s rationale.

# **Appendix A**

| **LBE** | baseline value (medical expert) | **DR** | repetitive decelerations | **C** | calm vigilance |
| --- | --- | --- | --- | --- | --- |
| **LB** | baseline value (SisPorto) | **Width** | histogram width | **D** | active vigilance |
| **AC** | accelerations (SisPorto) | **Min** | low freq. of the histogram | **SH** | shift pattern (A or Susp with shifts) |
| **FM** | foetal movement (SisPorto) | **Max** | high freq. of the histogram | **AD** | accelerative/decelerative pattern (stress situation) |
| **UC** | uterine contractions (SisPorto) | **Nmax** | number of histogram peaks | **DE** | decelerative pattern (vagal stimulation) |
| **ASTV** | percentage of time with abnormal short term variability (SisPorto) | **Nzeros** | number of histogram zeros | **LD** | largely decelerative pattern |
| **mSTV** | mean value of short term variability (SisPorto) | **Mean** | histogram mean | **FS** | flat-sinusoidal pattern (pathological state) |
| **ALTV** | percentage of time with abnormal long term variability (SisPorto) | **Median** | histogram median | **SUSP** | suspect pattern |
| **mLTV** | mean value of long term variability (SisPorto) | **Variance** | histogram variance | **CLASS** | Class code (1 to 10) for classes A to SUSP |
| **DL** | light decelerations | **Tendency** | histogram tendency: -1=left assymetric; 0=symmetric; 1=right assymetric | **NSP** | Normal=1; Suspect=2; Pathologic=3 |
| **DS** | severe decelerations | **A** | calm sleep | **SMOTE** | Synthetic Minority Over-sampling Technique |
| **DP** | prolongued decelerations | **B** | REM sleep |  |  |

Figure 1: Legend of Variables

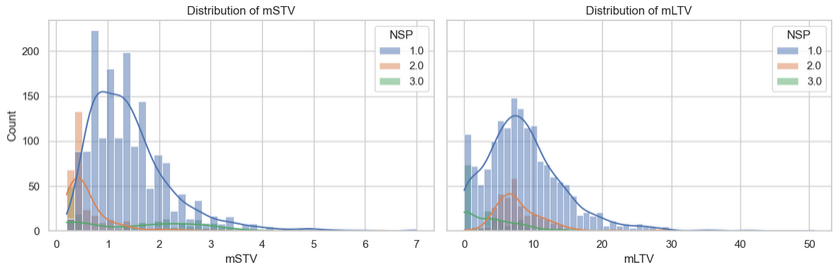


Figure 2: Histograms for mSTV (mean value of short term variability) and mLTV (mean value of long term variability)

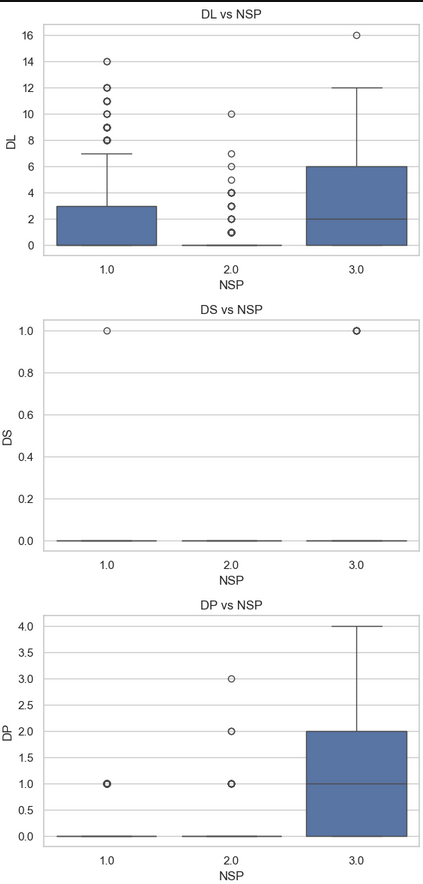


Figure 3: Boxplots by class (DL vs NSP, DS vs NSP, DP vs NSP)