**Data Analysis**

We chose to use joint plots of the various variables against NSP so that we can look at the correlation value between each pairing and find the variables with the highest correlation. After looking at the heatmaps, we concluded that the 5 most relevant variables to consider are Class, DP, ASTV , ALTV and Variance.

**Data Preprocessing**

Using the dataset provided, we dropped duplicated and unnamed columns, converted all numerical features to numerical types and removed label leakage columns. After doing a stratified 70:30 train-test split, we standardised using StandardScaler and computed class weights and sample weights to handle imbalance between classes.

**Model Testing**

Using the seven models provided from the workshop, we trained each model and evaluated them using Balanced Accuracy, F1 Macro and Prediction Time as the main metrics. Additionally, confusion matrices were analysed to identify False Positives, False Alarms, and Gray Areas. We also did some very light edits to some of the models to help them perform better without using up too much of our time.

From the above criteria, Neural Network (MLP) and k-Nearest Neighbours were the first 2 models we excluded. Their False Positive rates were too high to be considered. SVM still had comparatively high False Positive rates while Random Forest models were much slower (0.06s) compared to the other models (<0.01s). We decided to keep Decision Tree (for having the lowest False Positive rate) and Gradient Boosting (for its extremely low False Alarm rates) to do more detailed testing.

After additional testing, we chose Gradient Boosting as the final model for this datathon. Random Forest had lower False Positive rates but too much False Alarms that we were afraid it will erode clinician’s confidence.

This model is then constructed in a custom ThresholdClassifier to adjust how the %confidence for a sample to be placed at a certain class. This classifier also included a reject margin that reassigns uncertain samples to class 2 instead of risking a wrong class 1 or class 3 classification. By adding the reject margin, the model actually did slightly worse producing more false alarms (pred = 2, true = 1) but we felt this was important to keep false positives low.

The final Gradient Boosting model together with the ThresholdClassifier pipeline provided the best trade-off between having very low false positive rate yet low enough false alarms to ensure trust in clinicians. This model is therefore suitable in helping obstetric practitioners in identifying fatal cases at risk without generating excessive false alarms.