Data Cleaning

The dataset was carefully cleaned to ensure structural integrity, medical realism, and fairness in modeling. The engineered "Data" sheet was imported and standardized by removing redundant "Unnamed" columns, blank rows, and duplicate headers, while aligning feature names with CTG terminology (eg MSTV → mSTV, MLTV → mLTV). The target variable NSP (Normal = 1, Suspect = 2, Pathologic = 3) was defined, and label-leakage columns (CLASS, A–SUSP) were removed to prevent bias. Descriptive statistics and .info() checks confirmed numeric consistency, while a bar chart of NSP distribution revealed severe class imbalance (≈ 70% Normal, 20% Suspect, 10% Pathologic), justifying the later use of Balanced Accuracy, Macro F1, class weighting, and SMOTE. Features were grouped by medical relevance—physiological, variability, histogram, timing, and baseline—to improve interpretability. Median imputation was applied to handle rare missing values (< 1%), chosen for its robustness against skewed distributions such as ASTV and DP. Boxplots were used to visualise outliers, showing that high DP and abnormal ASTV values strongly correlated with Pathologic cases, while higher LB (baseline heart rate) and AC (accelerations) corresponded to Normal outcomes—aligning with clinical theory. No duplicate rows were found, and outliers were retained as they represent true distress events rather than noise. Overall, the data-cleaning process produced a fully numeric, medically interpretable, and balanced dataset supported by visual validation through descriptive plots, forming a robust foundation for subsequent correlation and predictive modeling.

Data Exploration

Exploratory analysis was performed to gain insights into the relationships and distributions within the CTG dataset. The correlation heatmap revealed strong positive relationships between features related to fetal heart rate variability such as mean value of short-term variability (MSTV) and long-term variability (MLTV), as well as accelerations and fetal movement. These associations suggested that higher fetal activity is often linked to greater variability in heart rate (an indicator of healthy fetal conditions). Distribution plots showed that the most physiological measures were right-skewed, indicating that extreme high values occurred less frequently. Box plots highlighted the presence of several outliers, particularly in variability-related features, which may correspond to abnormal fetal states. Overall, these exploratory findings provided valuable understanding of feature behaviour and relationships, guiding our predictive model better.

Predictive Model

We developed a prototype machine learning model that classifies the state of fetal well-being based on the cardiotocography (CTG) features and data provided. The model predicts using the three classes of Normal, Suspect and Pathologic (NSP), in order to allow for early detection of fetal distress during a mother's labour. For the purpose of training the model, we generated a dataset of 5000 samples using clinical interpretation rules. The four key features that we narrowed down to simulate were Baseline Fetal Heart Rate (LB), Variability, Accelerations and Decelerations. A rule-based function we utilised in our code would emulate standard guidelines such as absent variability with no accelerations, and late or prolonged decelerations would be labelled as "Pathologic".

Earlier, we also trained a random Forest classifier within a scikit-learn pipeline. The categorical variables of the dataset were one-hot encoded and the dataset split into training and test sets with stratified sampling. Thus the predictive model's performance could be evaluated using these classification metrics and a confusion matrix, proving the model has effectively learned the embedded clinical decision rules to diagnose fetal distress.

Lastly, we implemented a manual input interface, allowing users to enter the 4 key features selected as the CTG parameters for each individual case. The trained model would then output the predicted NSP class as well as its respective associated probabilities. Therefore, this demonstrates how this could evolve into a system that could be used as a real-time decision support tool in a clinical setting.