**TM59: Marcus Ho Geng Yip, Seth Yew Kuan Yang, Pham Hong Phuc**

**1. Dataset and Python libraries used.**

The dataset was taken from the Kaggle source provided. We took the ‘raw data’ sheets, and exported it as a csv for ease of processing on Jupyter Notebook and pandas for further processing. The csv file is named *CTG\_rawdata.csv* and the working notebook is named *4 notebook.ipynb*.

**2. Cleaning + Feature engineering**

We cleaned the data by clearing any potential duplicates and dropping the unneeded class columns. We created new features called, ‘runtime’ and ‘X\_rate’, where X are the clinical features to be normalised by the given runtime of the data survery.

**3. Exploratory analysis**

Several features, including ASTV, MSTV, and ALTV, display partial separation between classes, while others show significant overlap, suggesting complex multivariate relationships. Many features are right-skewed and non-Gaussian, reinforcing the use of ensemble and tree-based classifiers that handle non-linear feature interactions without distributional assumptions. Overall, these visual insights support the motivation for applying SMOTE to address imbalance and for adopting robust ensemble models for classification. Shown below are some of the produced charts. More can be found in ‘4 notebook’.

|  |  |
| --- | --- |
| A graph showing distribution of nsp classes  AI-generated content may be incorrect. | A screenshot of a graph  AI-generated content may be incorrect. |

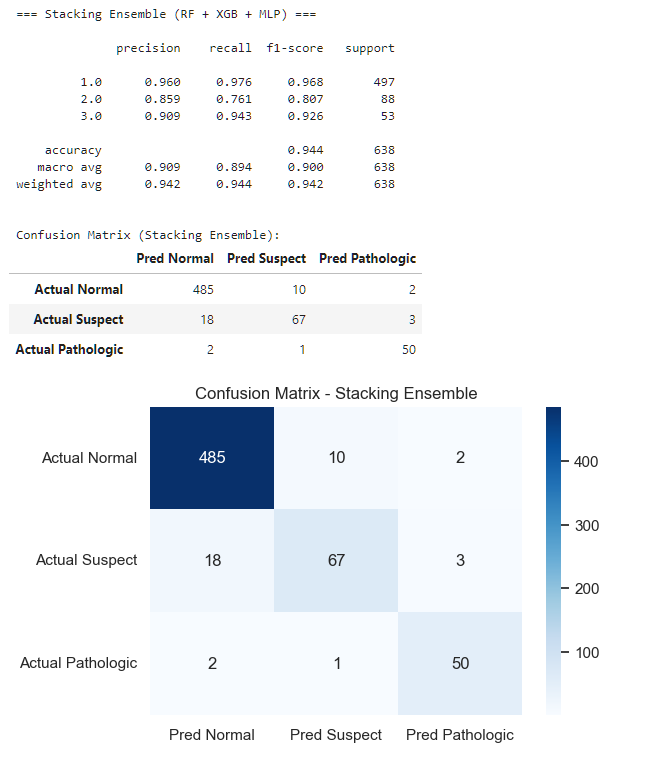
**4. Methodology**

Since this was a classification problem, we used random forest tree, XGBoost and Multi-Layered Perceptron (MPL). In fact, we created a stack ensemble of the 3, together with logistic regression to give a more robust model. The ensemble effectively balances linear interpretability and nonlinear flexibility, improving macro F1-score and reducing variance introduced by any single classifier.

Additionally, we also implemented Synthetic Minority Over-sampling Technique (SMOTE) is a method to address class-inbalances by generating synthetic datapoints via K-Nearest-Neighbour (KNN). We split the train-test data by 70/30.

**5. Results**

The stacked ensemble model, integrating Random Forest, XGBoost, and a Multilayer Perceptron (MLP), achieved excellent performance with an overall accuracy of 94.4% and a macro F1-score of 0.90. This demonstrates the ensemble’s ability to generalize effectively across all three fetal states—Normal, Suspect, and Pathologic. As shown in the confusion matrix, the model correctly classified the majority of Normal and Pathologic cases, with minor confusion occurring between the Normal and Suspect classes. This overlap is expected due to the subtle differences in fetal heart rate patterns between these two categories. The high recall and precision values indicate strong predictive reliability, particularly for detecting critical Pathologic cases, which are the most clinically significant. Compared to individual models, the stacked ensemble provides more balanced performance, leveraging the complementary strengths of tree-based learners and neural architectures to achieve both stability and adaptability in classification



**A graph with different colored bars

AI-generated content may be incorrect.**

**6. Conclusion**

Feature analysis using permutation importance revealed that **ASTV (Abnormal Short-Term Variability)** was the most dominant predictor of fetal health, followed closely by **AC\_rate (Acceleration Rate)**, **ALTV (Abnormal Long-Term Variability)**, and **MSTV (Mean Short-Term Variability)**. These features all describe aspects of fetal heart rate variability, a well-established indicator of fetal distress in cardiotocography studies. Other statistical descriptors, such as **H\_Mean**, **H\_Median**, and **H\_Variance**, also contributed meaningfully, reflecting how signal distribution characteristics influence classification outcomes. The prominence of variability-related features aligns strongly with medical understanding that reduced variability often signals compromised fetal oxygenation. Overall, the feature importance results validate the ensemble’s interpretability—confirming that the model not only performs accurately but also bases its predictions on physiologically meaningful variables, enhancing its potential clinical applicability.