**Datathon Academic Report | TM-75**

Model Design

|  |  |
| --- | --- |
| Model Description | The final model is an XGBoost classifier optimized to analyze cardiotocography (CTG) recordings for the early detection of fetal distress. XGBoost employs a boosting technique, meaning it builds decision trees sequentially rather than in parallel. Each new tree is trained to correct the errors made by the previous ones, allowing the model to progressively improve its predictions. The model optimises a softmax-based objective function using gradient boosting, which enables effective multi-class classification — hence the term “gradient boosting.” |
| Scoring | The following is the evaluation of the model with the current train-test split:  A screenshot of a computer screen  AI-generated content may be incorrect.  The model has also determined the following feature importances based on the gain for each feature:  A graph with numbers and a bar  AI-generated content may be incorrect. |

Rationale

|  |  |
| --- | --- |
| Justification | As the objective of the model was to classify cases, a 3-class XGBoost Classifier was deemed to be the optimal model over other models to categorise cases into either Normal, Suspect or Pathological. For hyperparameter tuning, we used random search rather than grid search as it is less computationally intensive. Subsequently, to choose the best hyperparameter combination, we evaluated the models using a weighted f1-score to balance the precision and recall. |

Data Processing Pipeline

|  |  |
| --- | --- |
| Cleaning | To begin the data processing pipeline, it is important to first clean the data – this entails removing data points that were unnecessary or interdependent. This includes DR, mode and mean. This reduces overweighting of less relevant data. Additionally, rows 2126-2128 are also of NaN values, which were removed in the process. The raw data from the first seven columns were also removed, to utilise only scaled data for the analysis. Columns A-SUSP were also removed as the model used a 3-class NSP classification. Min and Max were also dropped as their value was inherently embedded within Width. |
| Transformation | Once data has been cleaned, some data that are in percentages are scaled down to 0 to 1, such as the ASTV and ALTV columns. |
| Preparation | Data was train-test split into an 80-20 ratio for the model. |