Capstone Project

Fetal Risk Detection

with

Machine Learning

Drew Haszard
May 2025

Problem statement

Fetal complications remain one of the leading causes of perinatal mortality around the world. Despite advancements in technology, many hospitals still rely on the manual interpretation of cardiotocograms (CTGs)—a process that can be slow, subjective, and inconsistent across clinicians.

This challenge calls for a solution that combines medical knowledge with the precision of modern data science. By developing an automated, interpretable machine learning model, we aim to support healthcare professionals in making faster, more accurate decisions when it matters most.

This problem has been addressed by the **SisPorto 2.0** originally developed in the year 2000, a rule-based system used in the **Omniview** platform to automatically analyze CTG data. It was designed to help clinicians by offering real-time feedback on fetal heart rate and contractions, aiming to reduce human error and improve consistency.

While the older model showed promise, it relies on fixed rules, which can limit flexibility and adaptation to different patients. It also requires good infrastructure, making it harder to use in lower-resource settings.

This project builds on that work, using machine learning to create a model that can learn from specific CTG data, adapt to more cases, and be more accessible — especially where experienced CTG interpreters may not be available.

World Statistics of Perinatal Mortality:

WHO

Paper Citation:

Sisporto 2.0 Automated Analysis of Cardiotocograms

Latest Sisporto Real-World Model:

Omniview Medical Software

Industry / Domain

This project sits within the **healthcare industry**, specifically in the domain of **maternal and fetal health monitoring**. Accurate interpretation of cardiotocography (CTG) is a critical part of intrapartum care, helping clinicians detect early signs of fetal distress during labour.

Current State:
While CTG is widely used, interpreting the readings remains subjective, time-consuming, and requires experience. Variability among clinicians leads to inconsistent decisions, which can affect outcomes. Startups and research groups are actively exploring AI-powered tools to bring consistency and speed to CTG analysis. However, many solutions are either too rigid (rule-based systems) or not yet trusted in real-world clinical settings due to explainability and bias concerns.

Industry Value Chain:
The CTG value chain involves:

Data acquisition (CTG machines) → **Interpretation (Clinicians or AI tools)** → **Clinical Decision-Making (Interventions, Monitoring)** → **Patient Outcomes**

This project contributes to the interpretation stage.

Key Concepts:
Key concepts include **fetal heart rate variability**, **contraction frequency**, **clinical decision support systems (CDSS)**, **AI in healthcare**, and **model interpretability and fairness**.

Stakeholders

Clinical Practitioners, Midwives, and Obstetricians are the primary stakeholders in this project. They are directly involved in monitoring fetal health during labor and delivery, interpreting CTG readings, and making critical decisions based on the results.

Why they care:

These stakeholders are responsible for the safety of both the mother and the fetus during labour. Accurate and timely interpretation of CTG data is crucial for identifying potential risks and making decisions that can improve patient outcomes.

Stakeholders' expectations:

Clinical Practitioners and Midwives expect a reliable, easy-to-use tool that helps them quickly and accurately classify fetal risk, reducing the reliance on manual CTG interpretation, particularly in busy or low-resource settings. They also expect the system to assist in decision-making by providing consistent and evidence-based recommendations.

Business question

How can we automate the classification of fetal risk levels from CTG readings to assist clinical practitioners in real-time?

Business Value:

Automating CTG interpretation improves clinical efficiency, reduces human error, and supports timely intervention—potentially lowering rates of fetal and maternal complications. Value can be seen in better outcomes, saved clinician time, and reduced training demands in low-resource settings.

Required Accuracy & Implications:

A high-performing model is essential, especially for identifying high-risk cases. We optimize for **macro F1-score**, which balances precision and recall equally across all classes. This is critical in a medical setting where **false negatives** (missing a high-risk case) can have severe consequences. The macro F1-score ensures that smaller, riskier classes—like High Risk—are not overshadowed by the majority class.

Data question

Can fetal risk levels be accurately predicted using statistical features extracted from CTG signals?

Data Required:

Labeled CTG data with summary statistics for features like fetal heart rate, and uterine contractions, along with classification fetal risk/health labels. This allows supervised learning to train and validate classification models.

Data

This dataset originates from the *SisPorto 2.0* system, developed around the year 2000 during its research phase.

Although the cited SisPorto 2.0 paper refers to a preliminary study using only 85 recordings, our dataset contains over 2,000 rows. This strongly suggests the data was sourced from a **broader database** of CTG recordings analyzed by the SisPorto system. While the original paper does not specify trace length, standard clinical practice supports the use of **20-minute CTG recordings**, making it reasonable to assume the majority of these traces follow that protocol.

This dataset originates from the *SisPorto 2.0* system, developed around the year 2000 during its research phase.

** Please reference jupyter notebook for feature definitions**

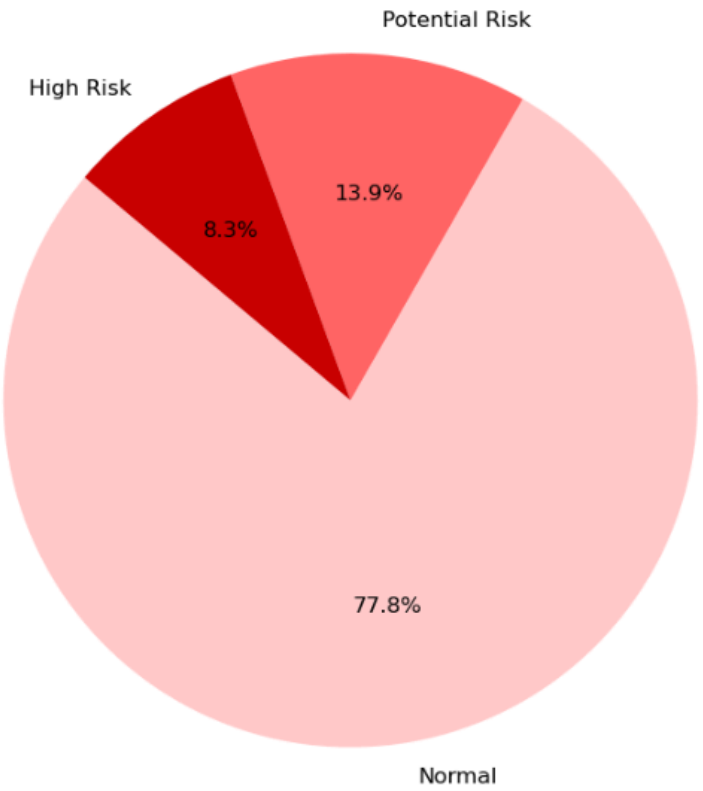
- **Source:** Kaggle, based on research using SisPorto 2.0 system (Kaggle Dataset)
- **Volume & Attributes:** 2,126 recordings (rows) with 21 numerical features capturing statistical summaries of fetal heart rate and uterine contractions, plus 1 labeled fetal health classification feature.
- **Reliability & Quality:** The data was clinically labeled by experts, but represents historical, research-phase technology and lacks modern clinical diversity.
- **No missing data**
- **No data leakage**
- Some features are **summary statistics** of the fetal heart rate over the whole monitoring period
- **Ongoing Availability:** The dataset is static and not updated, highlighting the need for more modern, real-time data in future research.

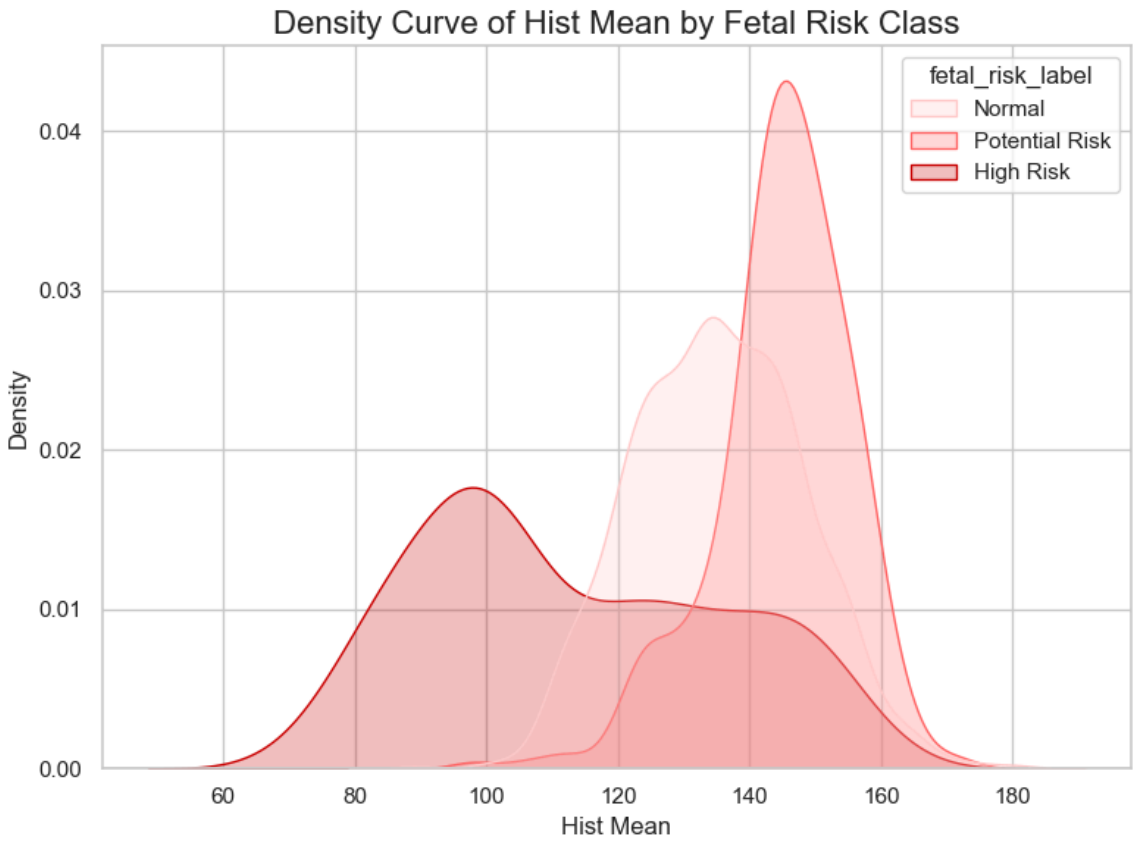
Data science process

Data analysis

The data analysis was performed using Python (Pandas, NumPy, Scikit-learn), and involved the following key steps:

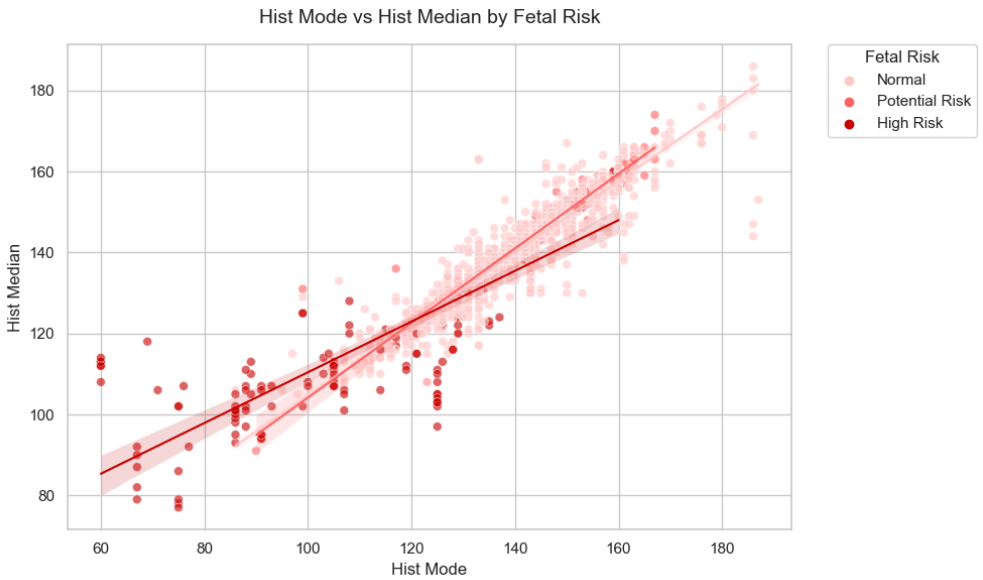
- **Loading the CSV** and inspecting structure
- **Checking for missing or duplicated values**
- **Renaming columns** for readability and consistency
- **Creating a new feature fetal_risk_label** (labelled version of fetal_risk)
- **Removing non-informative features**, such as binary columns and the text-based fetal_risk_label
- Performed **EDA on target variable** (Noticed class imbalance)





Plot above is the mean fetal heart rate (BPM) by Fetal Risk Class, we see that the higher risk class tends to have a lower average BPM and the potential risk has a higher range of values. This plots show that a higher BPM could indicate lower fetal risk

- Created a **Correlation Matrix**
- Performed more **EDA** and **plotted visualisations** to inspect features that had multicollinearity



The plot above shows a clear positive linear relationship between hist_mode and hist_median. As the median fetal heart rate (hist_median) increases, the mode of the heart rate (hist_mode) also tends to increase across all risk categories. This result aligns with logical expectations and supports the consistency of the data.

The multicollinearity noticed in this analysis was considered in future modelling, along with the class imbalance.

- **Fetal_risk_label feature was then dropped** to avoid data leakage and remove any worded features as numeric data was needed for modelling

Key findings from EDA:

- The target variable (fetal_risk) was **imbalanced**, with ~77% of cases labeled as Normal.
- **Correlation analysis** revealed moderate to strong relationships between variables such as mean_value_stv, percent_abnormal_ltv, and fetal_risk
- **Outliers** were present in some clinical variables, which informed the decision to use RobustScaler over StandardScaler.

Modelling

Overview
I implemented two distinct modelling pipelines to predict fetal risk using classification models. Each pipeline followed a consistent process of preprocessing, model training, and evaluation. The pipelines differ primarily in their use of hyperparameter tuning.

Models: Random Forest, XGBoost, Gradient Boosting

Class Imbalance
To handle class imbalance, I applied **sample weights** to the XGBoost and Gradient Boosting models and used **class_weight='balanced'** for the Random Forest model. This ensured that the models gave appropriate attention to underrepresented fetal risk categories without oversampling or synthetic data.

- Robust Scaler was used to handle the large amount of outliers in the data.
- Train/Test split 80/20
- Randomized Search was used for hyperparameter tuning
- Cross Validation was used for pipeline 2
- Macro f1 score was chosen as performance metric because it averages the f1-score from all classes - so all classes regardless of imbalance will be fairly included

Pipeline 1: Baseline Models (No Tuning)

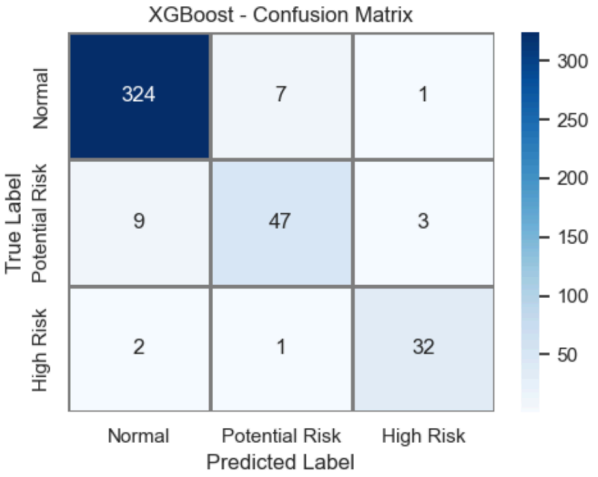
All three models showed signs of overfitting, with their training scores plateauing at 1.0 in the learning curves. This indicates that the models fit the training data perfectly but struggled to generalize to unseen data. Among them, Gradient Boosting showed the smallest gap between training and validation curves, suggesting slightly better generalization. The overfitting likely stems from the absence of hyperparameter tuning, particularly for controlling model complexity (e.g., tree depth). Despite this, XGBoost achieved the highest macro F1-score, demonstrating strong predictive performance — though its overfitting highlights the need for further tuning to improve robustness

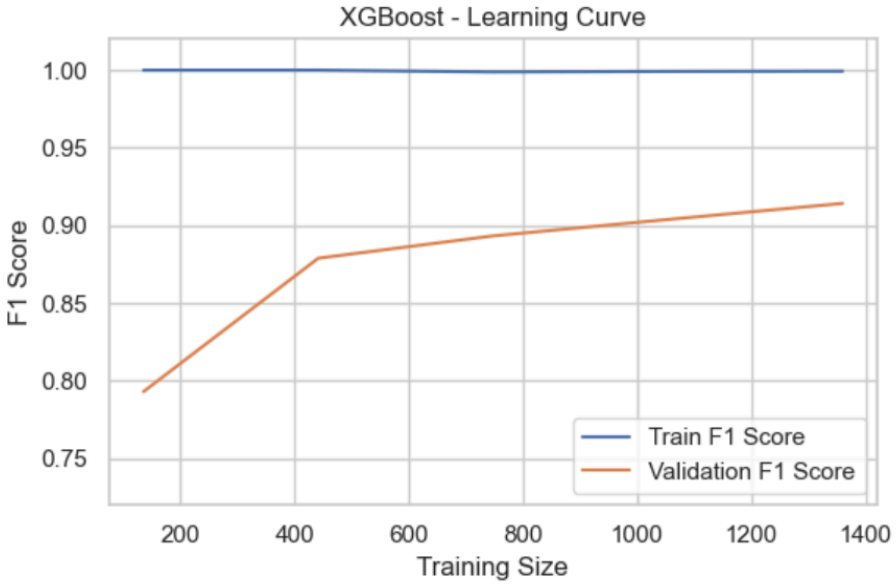
XGBoost Performance:

```
XGBoost
=======
Classification Report:
                precision    recall  f1-score

        Normal     0.9672    0.9759    0.9715
Potential Risk     0.8545    0.7966    0.8246
     High Risk     0.8889    0.9143    0.9014

      accuracy                         0.9460
     macro avg     0.9035    0.8956    0.8992
  weighted avg     0.9451    0.9460    0.9454
```





- **Macro F1-score: 0.8992** - Highest of all models

- ***Confusion Matrix:***
  - **Normal:** 324/332 correct - **97.6% recall**
  - **Potential Risk:** 47/59 correctly identified → **79.7% recall, best among all models;** means nearly 4 in 5 Potential Risk cases were correctly caught.
  - **High Risk:** 32/35 - **91.4% recall**

- ***Learning Curve:***
  - **Training line:** Flat at 1.0 - very high training performance, suggesting **overfitting**.
  - **Validation line:** Starts at ~0.80 and increases steadily → indicates the model continues to learn as more data is used, but **is overfitting**

- ***Precision & Recall Summary:***
  - Strong performance across all classes, especially **excellent recall on the Normal class**
  - High recall means **fewer false negatives** — critical for a risk prediction task
  - High precision ensures **predicted risks are rarely false alarms**

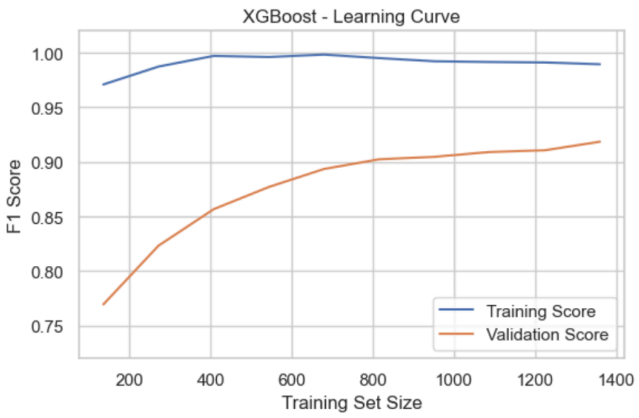
Pipeline 2: Baseline Models (Hyperparameter Tuning)

After hyperparameter tuning, **XGBoost was selected as the best overall model**. While Gradient Boosting achieved the highest macro F1-score, it showed greater overfitting on the learning curve. In contrast, XGBoost had slightly lower performance but generalized better, making it the more stable and reliable choice—especially important in clinical applications.

XGBoost Performance:

Classification Report:

	precision	recall	f1-score
Normal	0.9750	0.9398	0.9571
Potential Risk	0.7353	0.8475	0.7874
High Risk	0.8684	0.9429	0.9041
accuracy			0.9272
macro avg	0.8596	0.9100	0.8829
weighted avg	0.9330	0.9272	0.9292



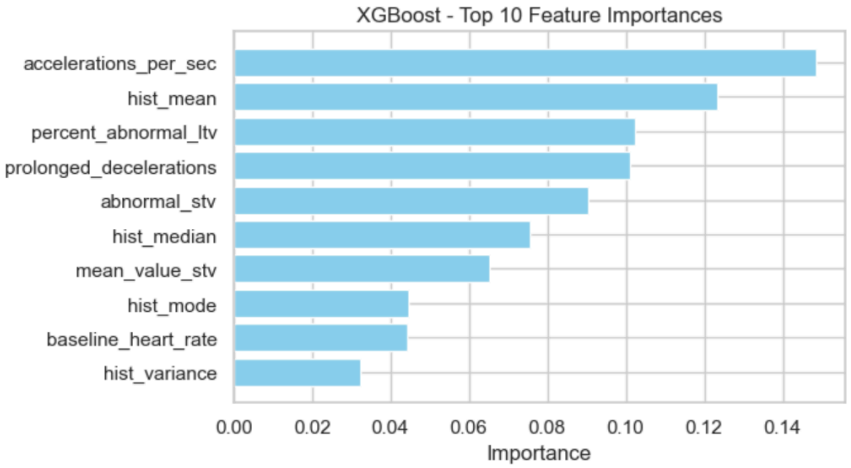
Macro F1-score: 0.899
Top Features: accelerations_per_sec, hist_mean, percent_abnormal_ltv

Confusion Matrix:

- **Normal:** Precision = 97.50%, Recall = 93.98%, F1-Score = 95.71%. Out of 332 predictions for Normal, 97.50% were correct, and 93.98% of actual Normal cases were correctly predicted.
- **Potential Risk:** Precision = 73.53%, Recall = 84.75%, F1-Score = 78.74%. Out of 59 predictions for Potential Risk, 73.53% were correct, and 84.75% of actual Potential Risk cases were correctly predicted.
- **High Risk:** Precision = 86.84%, Recall = 94.29%, F1-Score = 90.41%. Out of 35 predictions for High Risk, 86.84% were correct, and 94.29% of actual High Risk cases were correctly predicted.

Learning Curve:

- **Training line:** Slightly improved, showing less overfitting compared to Pipeline 1. The model still memorizes the training data to some degree, but it generalizes better than before.
- **Validation line:** The cross-validation line starts at 0.8 and increases, showing that the model generalizes well but doesn't match the training performance exactly.




```
Best Hyperparameters:
  classifier__subsample: 0.7
  classifier__reg_lambda: 0.5
  classifier__reg_alpha: 0.5
  classifier__n_estimators: 500
  classifier__min_child_weight: 3
  classifier__max_depth: 2
  classifier__learning_rate: 0.1
  classifier__gamma: 0.1
  classifier__colsample_bytree: 0.6
```

- Model Pipeline 2 using randomized search and cross validation = 5 took around 5 minutes to run
- Features were selected by the specific model
- XGBoost was the best performer for both models
- All models were over-fitting, this is most likely because there is such a small amount of data to work with so the models are memorizing the training data and noise too well.

Outcomes

- All models were over-fitting, this is most likely because there is such a small amount of data to work with so the models are memorizing the training data and noise too well.
- Having extra real world CTG data would be ideal for this model to learn the training data better
- There are a few redundant features that have multicollinearity, removing these may improve model performance
- Using less complex models may also simplify the machine learning algorithms and intern generalize better

After tuning all models using RandomizedSearch, XGBoost emerged as the best-balanced model overall.

Although Gradient Boosting delivered the highest macro F1 score at 89%, its learning curve revealed clear signs of overfitting. It performed almost perfectly on training data but didn't generalize well to new cases.

In contrast, XGBoost scored slightly lower at 88% but demonstrated more stable learning and better generalization. That reliability made it the stronger and safer choice—especially in a clinical setting where new data will be added over time so adaptability is crucial.

So, while Gradient Boosting initially appeared to be the top performer, XGBoost ultimately proved to be the more dependable model

So now we have a solution - A complete ML pipeline that splits and scales real CTG data, tunes and evaluates multiple models, and we've selected the best performer

Implementation

While the model can start as a web-based tool for real-time CTG interpretation in hospitals, it could also be adapted to run offline on solar-powered tablets or smartphones connected to affordable CTG devices (like Dopplers or basic cardiotocographs), improving diagnostic accuracy and supporting fetal risk detection in clinics without electricity or internet access.

Next, collaborating with midwives and clinical staff is essential to validate the model's predictions in real-world environments.

I'd also like to explore more advanced hyperparameter tuning with GridSearch, to push performance even further.

And finally, the model would benefit from more real CTG data—especially from different populations—to improve generalizability and reduce bias

Data answer

The data question — *Can fetal risk be predicted using cardiotocography (CTG) data?* — was answered satisfactorily. Confidence in the data answer is **moderate to high**, based on strong model performance. However, signs of overfitting were observed, indicating room for improvement in the model's ability to generalize to new, unseen data.

Business answer

The project successfully addressed the business question by demonstrating that Machine Learning can help assess fetal risk using CTG data. The model outputs show promising potential to support clinical decision-making, particularly in early risk detection. Confidence in the business outcome is **reasonably high**, though further improvements in model generalization and real-world testing are needed to build full trust in deployment.

Response to stakeholders

The results show that Machine Learning can provide valuable decision support during labor by flagging potential fetal risks from CTG data. I recommend piloting the model in environments with stable infrastructure first, then exploring offline deployment for rural clinics. Further model refinement and real-world validation are essential before clinical adoption, but this project highlights a scalable path toward safer, more consistent CTG interpretation.

End-to-end solution

The end-to-end solution involves the following steps:

1. **Data Collection & Preprocessing**: Raw CTG data is collected from fetal monitoring devices, then preprocessed to handle missing values, standardize features, and

address class imbalance using sample weights and balanced class weights.

2. **Model Training**: Two model pipelines were created. The first pipeline utilized Random Forest, Gradient Boosting, and XGBoost without hyperparameter tuning. The second pipeline focused on optimizing hyperparameters with RandomizedSearchCV to maximize the F1-score.

3. **Model Evaluation**: Models were evaluated on key performance metrics like F1-score, accuracy, and confusion matrix, with XGBoost showing overall strong performance but facing challenges with overfitting and generalization.

4. **Deployment**: The model is designed for both real-time and offline deployment, with an initial web-based tool for hospitals with stable internet. For low-resource settings, the model could be integrated into portable CTG devices or mobile applications, with offline capabilities to support clinics without electricity or internet access.

5. **Feedback & Improvement**: Continuous monitoring and feedback from healthcare professionals will guide further model refinement and validation to ensure accuracy and reliability for clinical decision-making.