# Predicting **Prenatal Health** with Machine Learning

Ironhack final project - Alessia

# Introduction

- In 2015, 45% of under-5 mortality occurred during the neonatal period, primarily due to birth complications (35%), intrapartum events (25%), and infections like sepsis or meningitis (15%) *Liu L. et al. Lancet (2016)*

- Cardiotocography (**CTG**) is used to assess **fetal well-being** during pregnancy. It helps identify complications and ensures timely medical intervention to protect both mother and baby

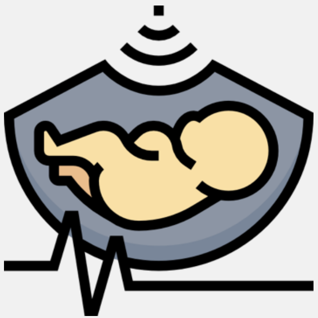
- CTGs measure fetal heart rate (FHR), variability, accelerations, and decelerations and are classified as:
  - Normal
  - Suspect
  - Pathological

# Objective

Analysing CTGs and drawing conclusions is challenging, especially in underdeveloped countries due to a shortage of skilled medical professionals

**AIM:**

develop a **machine learning model** to accurately **predict high risk fetuses** based on CTG exams results



# Data Overview

## Dataset

**Fetal Health Classification (Kaggle):**

- 2126 Rows and 22 columns
- No NaN
- 13 Duplicated values (dropped)

- All columns are Floats, including “Fetal Health”

## Target

**“Fetal Health”:**

- 1 = Normal
- 2 = Suspect
- 3 = Pathological

→ **Classification Problem**

*Ayres de Campos et al. (2000) J Matern Fetal Med 5:311-318*

# Project Workflow

### 1 — Data exploration & cleaning

- Removing duplicated values
- Conversion of target col into N, S, P for better visualization

```
df.shape
(2113, 22)
```

### 2 — EDA

- Target and Features distributions
- ANOVA testing
- Feature importance

### 3 — Tableau visualization

https://public.tableau.com/app/profile/urzi.alessia/viz/final_project_Ironhack/Dashboard12?publish=yes

### 4 — Pre-modeling

- Resampling
- Feature engineering
- Data Transformation

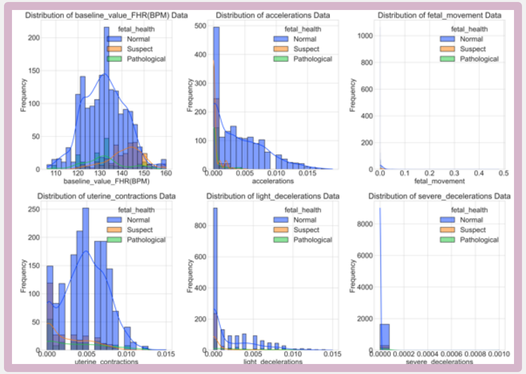
### 5 — Machine Learning

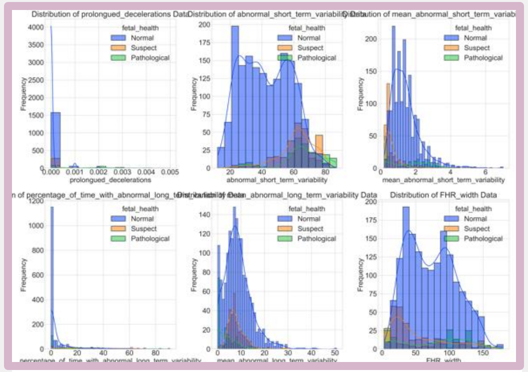
Classification models: KNN, AdABoost, RF, GradientBoost, XGBoost

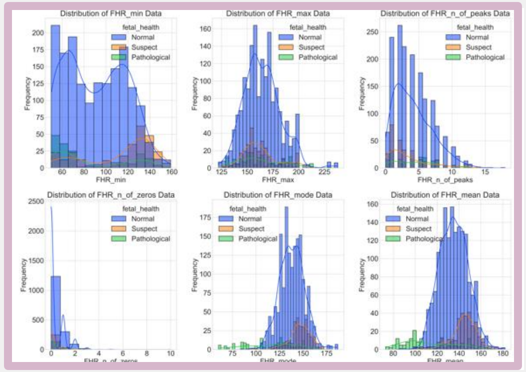
# Tableau visualization

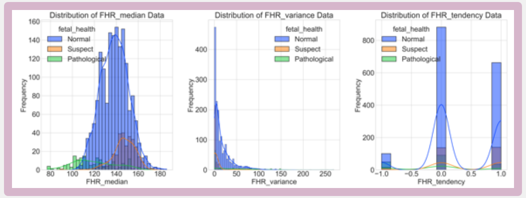
- Tableau Dashboard: Link

# Features Distribution




- Most Features follow a normal distribution pattern (FHR max, FHR median, FHR mean, FHR mode, base FHR)
- However, some Features are highly skewed (fetal movements, accelerations, decelerations, FHR variance)

# Machine Learning process

### Imbalance

**SMOTE** was used to resample the target column


### Transformation

**Robust Scaler**:

- Presence of outliers
- Skewed distributions

### Feature engineering

Conversion of target into numeric values with **Label Encoder (XGBoost)**, then decodified for interpretation

# Machine Learning Models

- A GridSearch was run to identify the best parameters for KNN, AdaBoost, GradientBoosting, RF, and XGBoost
- The best parameters were then applied to train each Model

| Model | Avg. Accuracy (weighted) | Avg. Precision (weighted) | Avg. Recall (weighted) | Avg. F1 Score (weighted) | Avg. Cohen's Kappa |
|---|---|---|---|---|---|
| AdaBoost | 0.90 | 0.93 | 0.90 | 0.91 | 0.75 |
| K-Nearest Neighbors | 0.92 | 0.93 | 0.92 | 0.92 | 0.80 |
| Gradient Boosting | 0.96 | 0.96 | 0.96 | 0.96 | 0.88 |
| Random Forest | 0.96 | 0.96 | 0.96 | 0.96 | 0.88 |
| **XGBoost** | **0.97** | **0.97** | **0.97** | **0.97** | **0.91** |

→ XGBoost algorithm does not require a normal distribution and is not affected by outliers and class imbalance

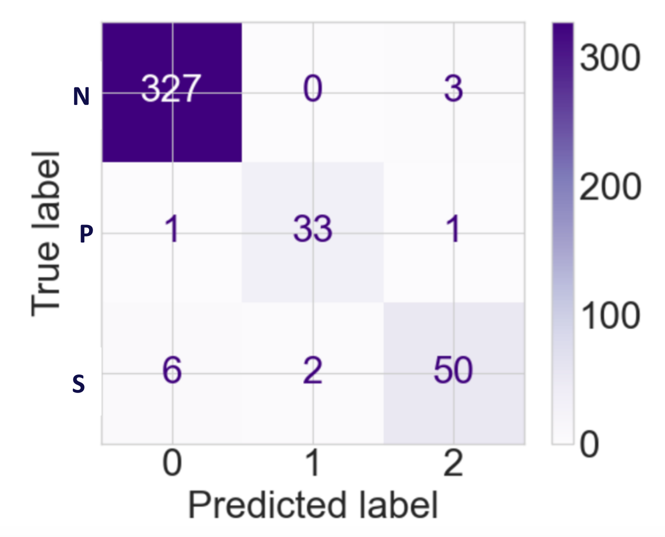
# Using XGBoost to predict fetal health

## XGB Classification report

| | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| Normal | 0.98 | 0.99 | 0.98 | 330 |
| Pathological | 0.94 | 0.94 | 0.94 | 35 |
| Suspect | 0.93 | 0.86 | 0.89 | 58 |
| Weighted Avg | **0.97** | **0.97** | **0.97** | |

| | | | | |
|---|---|---|---|---|
| Accuracy | 0.97 | | | |

## Confusion Matrix

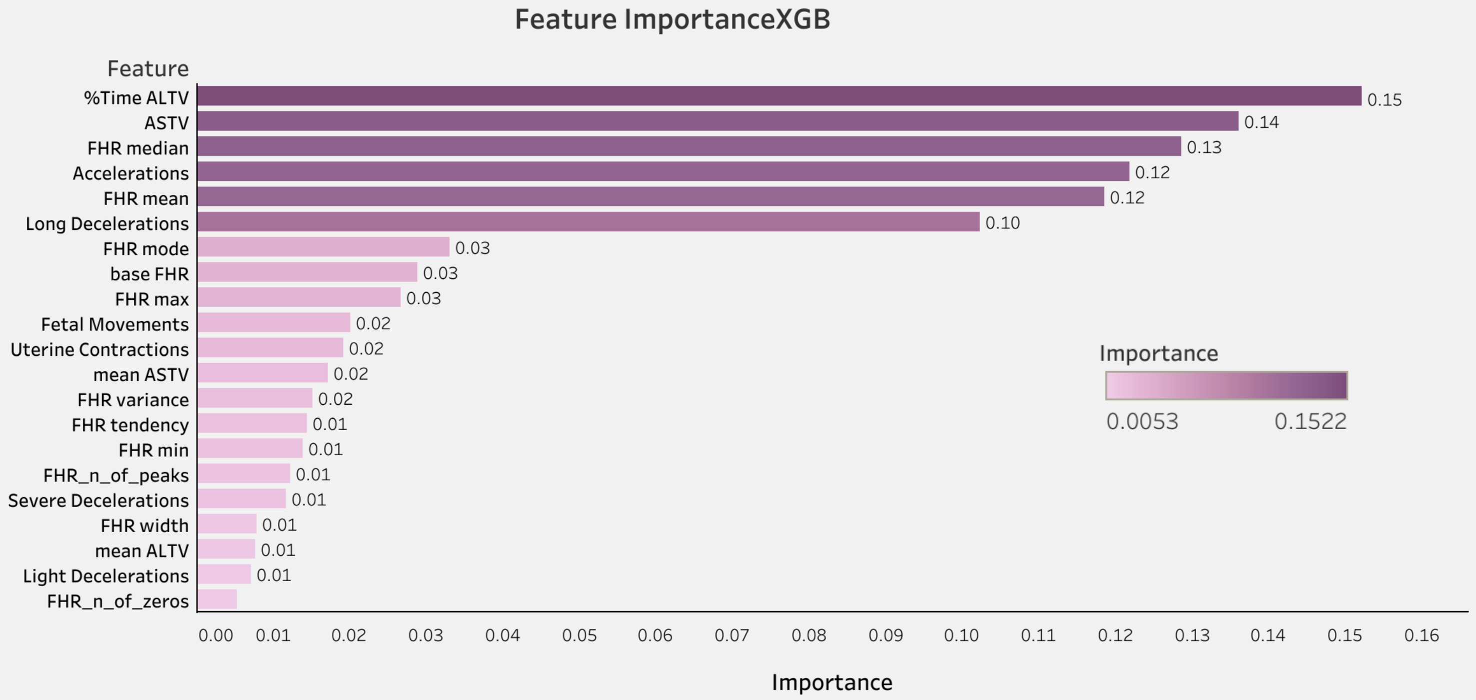


# XGBoost Train vs Test

| Metric | Avg. Accuracy (weighted) | Avg. Precision (weighted) | Avg. Recall (weighted) | Avg. F1 Score (weighted) | Avg. Cohen's Kappa |
|---|---|---|---|---|---|
| Training Set | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 |
| Test set | **0.97** | **0.97** | **0.97** | **0.97** | **0.91** |

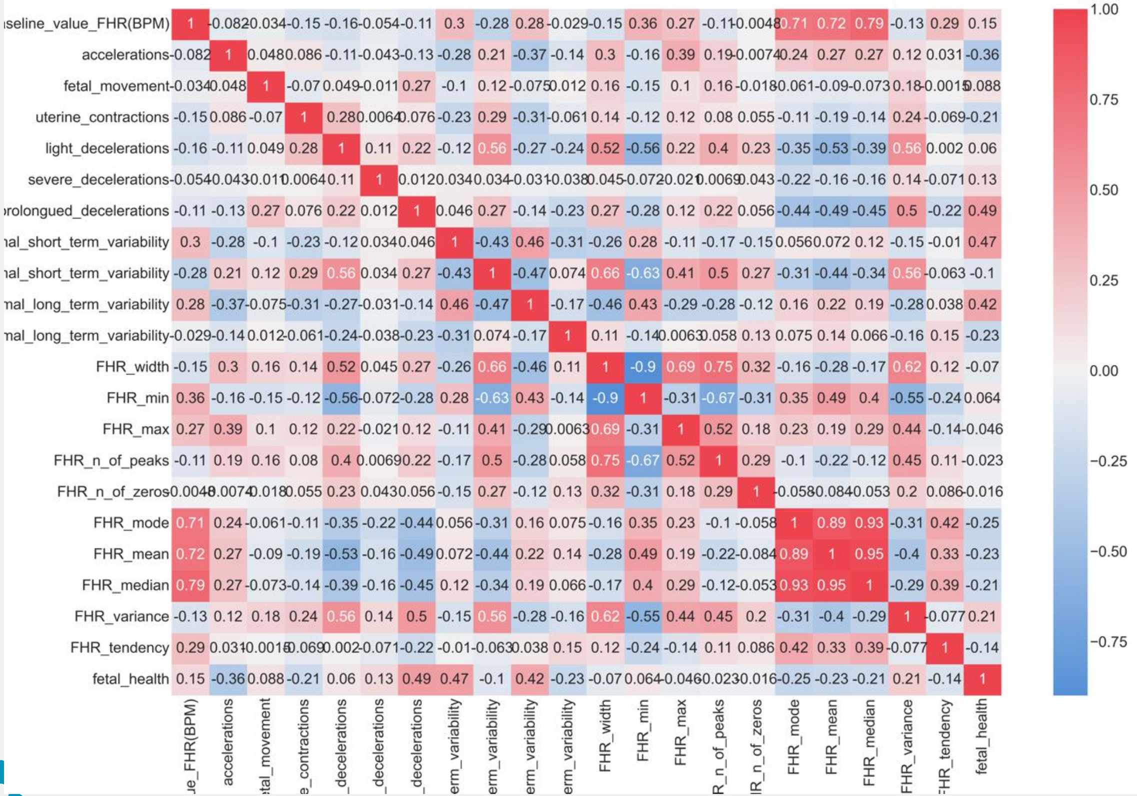
- XGBoost Model is generalizing well on both training and test sets

# XGBoost Feature Importance



# Features Correlation

## Correlation Matrix




**HIGHLY CORRELATED FEATURES (> 0.75):**

- baseline_value_FHR(BPM) and FHR_median

  FHR_width and FHR_min
- FHR_min and FHR_width
- FHR_mode and FHR_mean
- FHR_mode and FHR_median
- FHR_mean and FHR_mode
- FHR_mean and FHR_median
- FHR_median and baseline_value_FHR(BPM)
- FHR_median and FHR_mode
- FHR_median and FHR_mean

Highly correlated features could be removed to avoid redundancy

# Future Improvements

## Feature importance

Keeping only most important Features

## Feature correlation

Removing highly correlated features to reduce redundancy

## Re-Train XGBoost

Re-train XGB Model with these implementations to further improve performance and clarity

# Real Life application

## Streamlit App demo: FetalHealth

This application is designed to assist in identifying high-risk fetuses, even without trained medical professionals.

By inputting values from a CTG examination below, you can obtain predictions regarding fetal health status.

Link to App:

https://aleurzi-final-project-ironhack-streamlitapp-ln2uq4.streamlit.app/

# Thank you!

- Kaggle Dataset: Link
- Tableau Dashboard: Link
- Streamlit App: Link

*Ayres de Campos et al. (2000) J Matern Fetal Med 5:311-318*