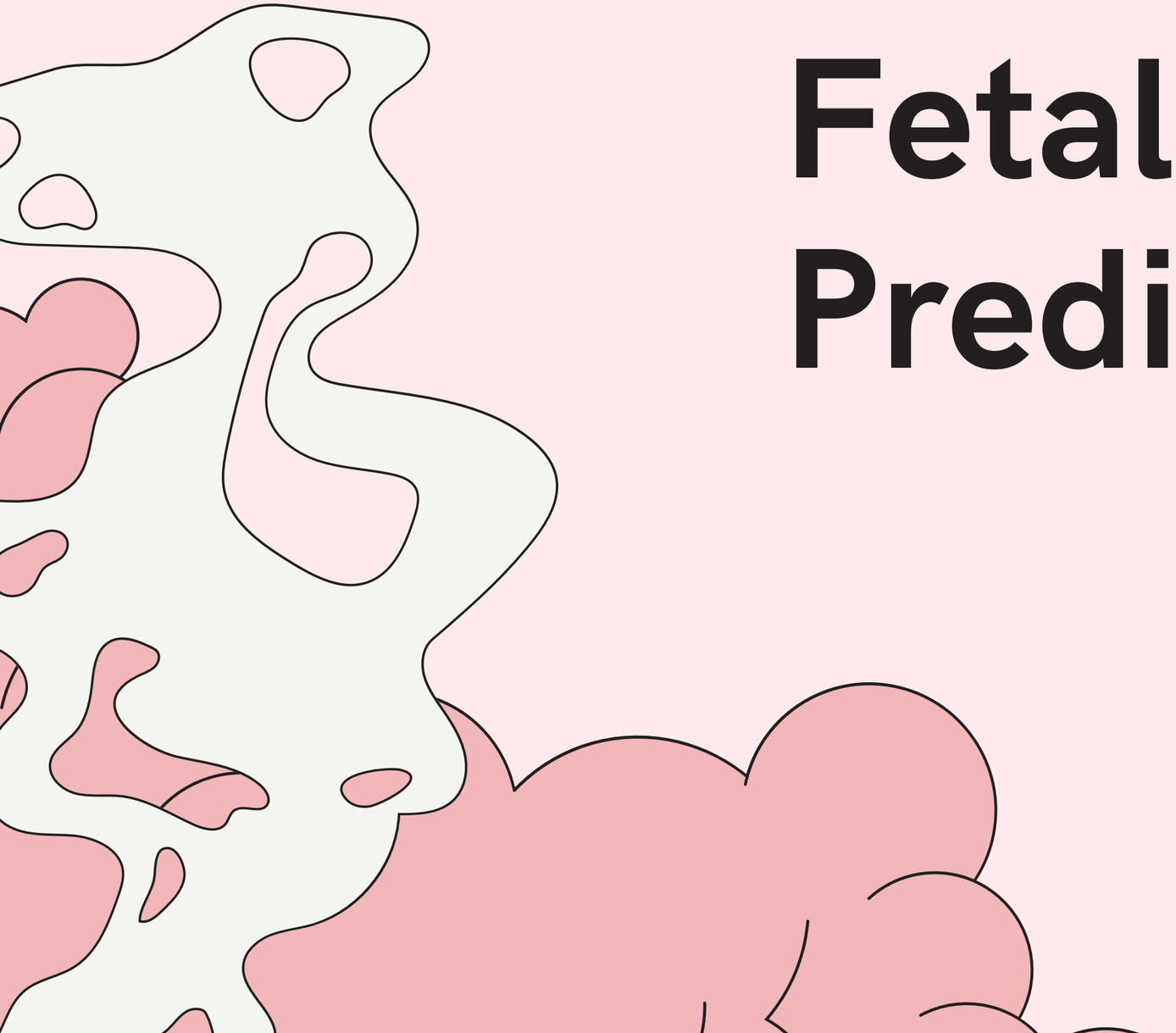
ibimbing



Fetal Health Prediction

Azri Ahza Ihtifazuddin

print(”Hello World!”)

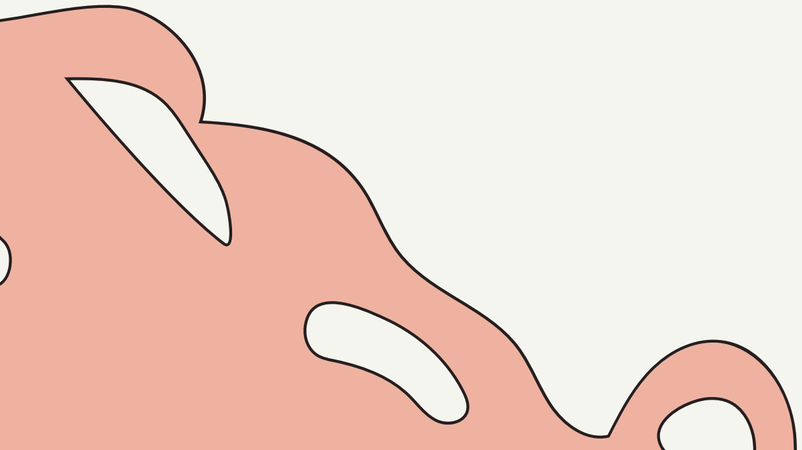
Azri Ahza Ihtifazuddin

Data Science Enthusiast

- A **Career switcher** with background **bachelor of sains**
- **Data scientist internship** at Bukit Vista
- Data science bootcamp at dibimbing.id
- Having experience as **internal auditor production** (Adidas Factory)

my-full-cv

My Data Projects



Fetal Health Prediction

full-article

Proyek ini memprediksi status kesehatan janin dari data kardiotokografi sehingga status kesehatan janin dapat diketahui lebih mudah dan cepat.

Data Scraping & EDA Bukit Vista

full-article

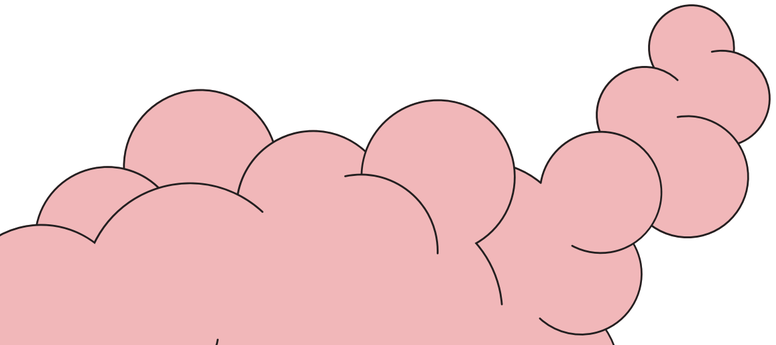
I scraped data from the Bukit Vista website. Data analysis from the point of view of a sales marketer, property reviewer, and customer.

Bee Cycle Dashboard using Tableau

full-article

Cost Dashboard which aims to summarize cost information for each product and type of expenditure from the company as a reference tool in expenditure analysis by management.

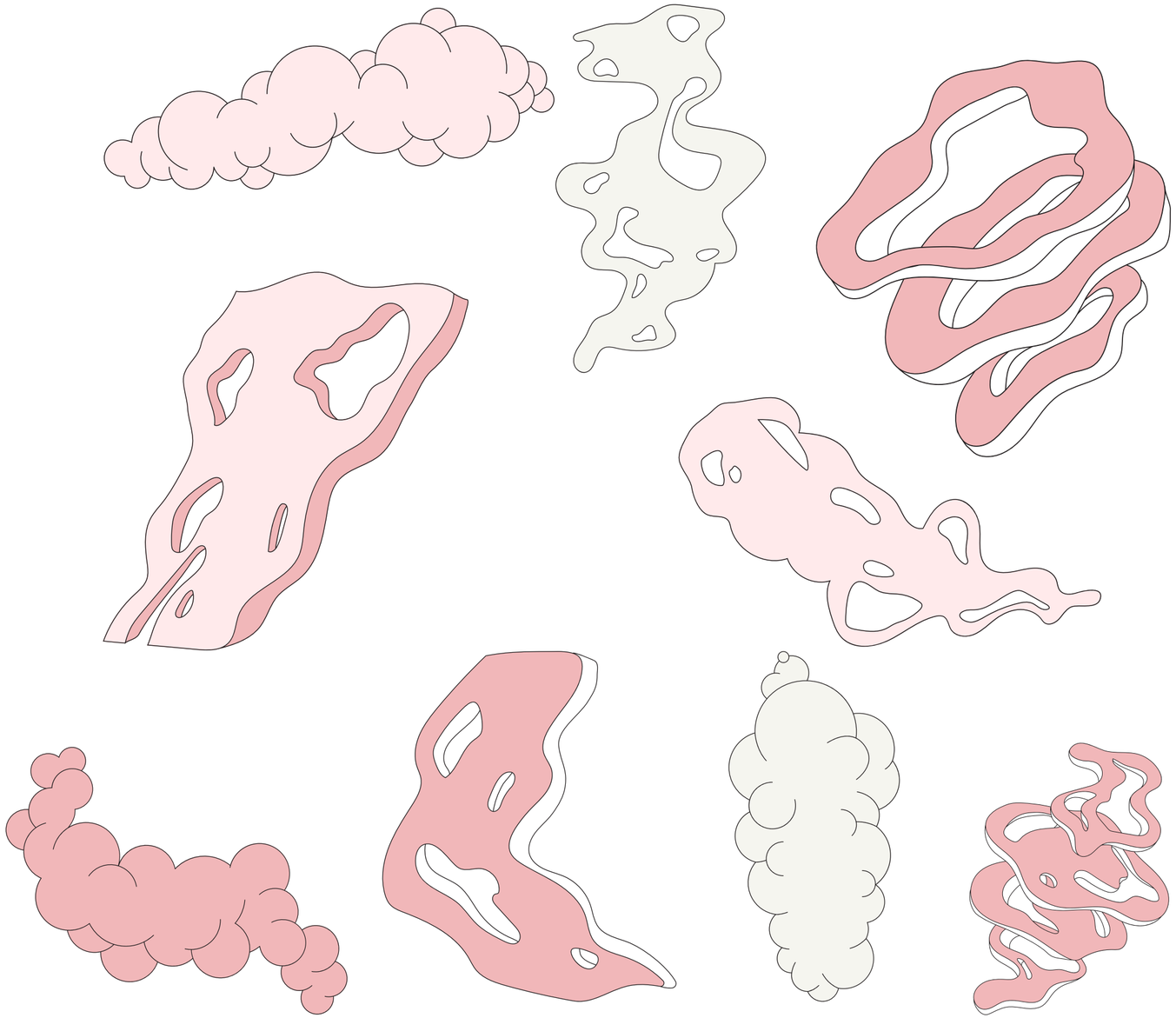
Table of Content



1

Overview & Business problem

Understanding Infant Mortality and Cardiotocography (CTG)



Infant death

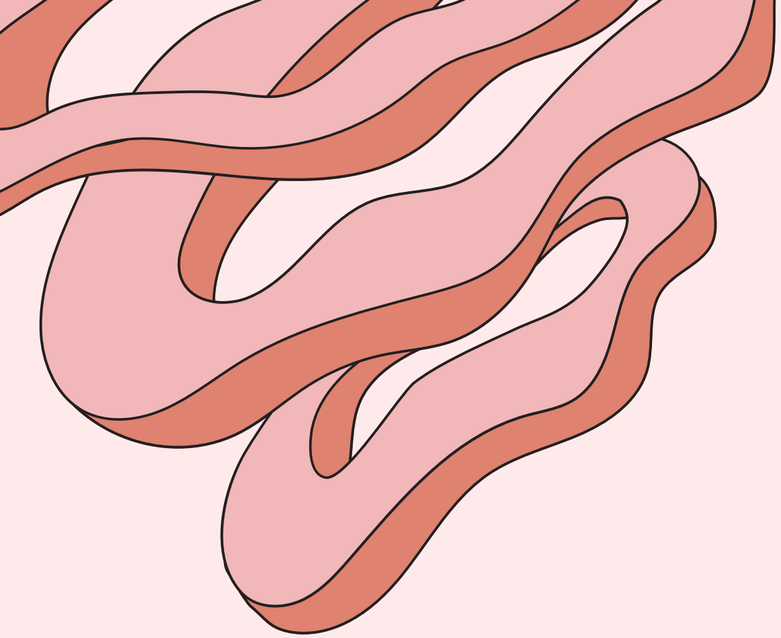
The infant mortality rate is 24 per 1,000

If 100 people give birth, between 2 and 3 die

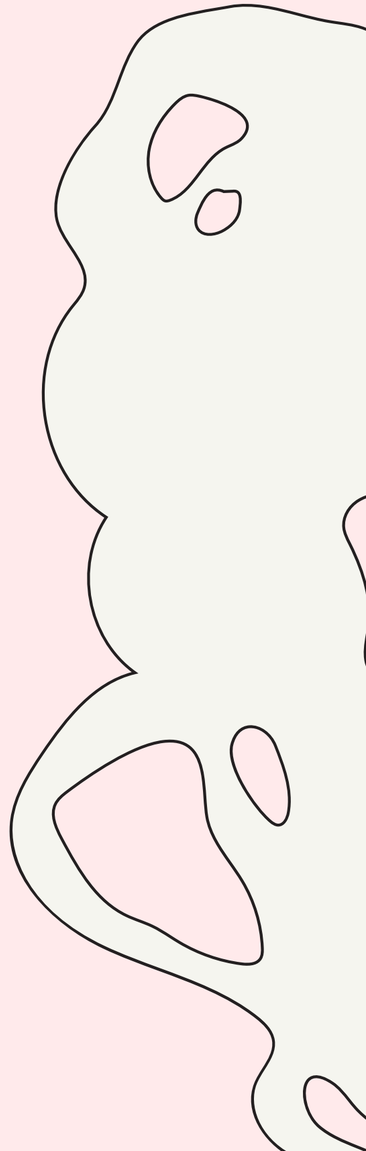
The health of pregnant women and their fetuses is interrelated

Unhealthy fetus is at risk of death during childbirth

Source: Badan Koordinasi Keluarga Berencana Nasional(BKKBN)



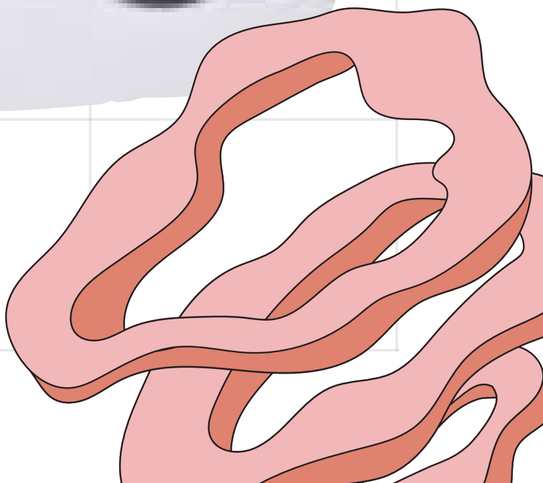
‼ The condition of the fetus is unknown

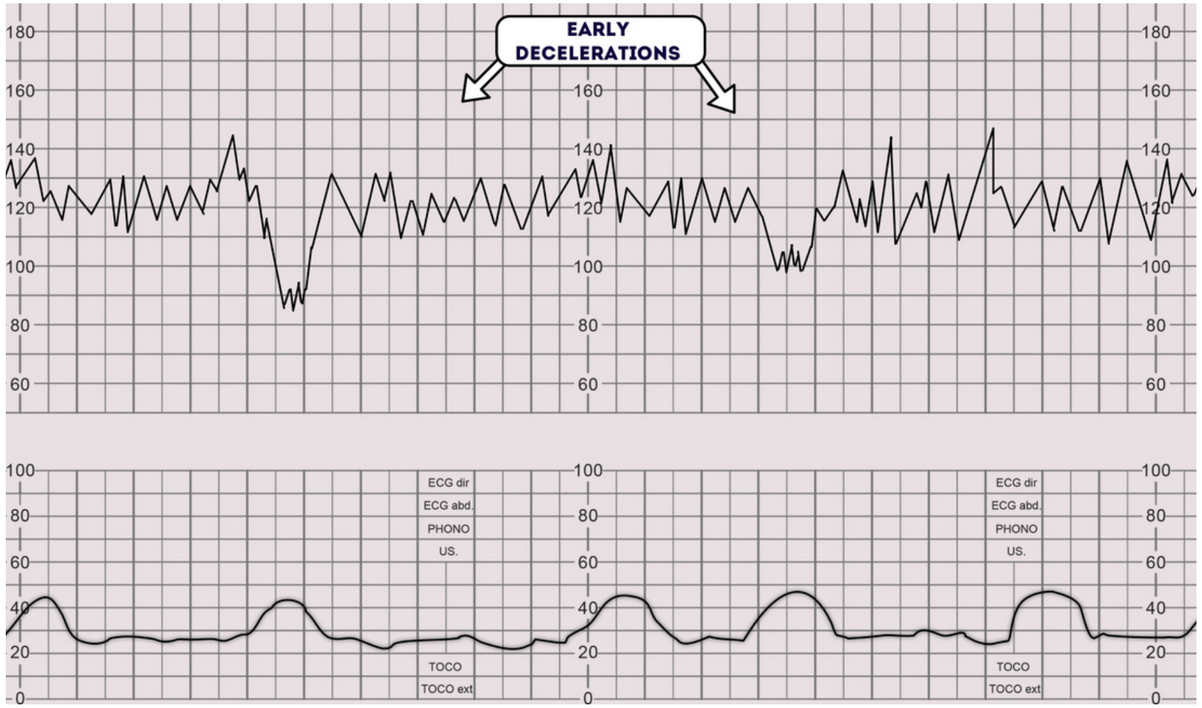


Cardiotocography (CTG) Method

Method is the process of **examining the health of the fetus** in pregnant women.

Prenatal monitoring using two CTG signals, namely, **fetal heart rate** and **uterine contractions**.



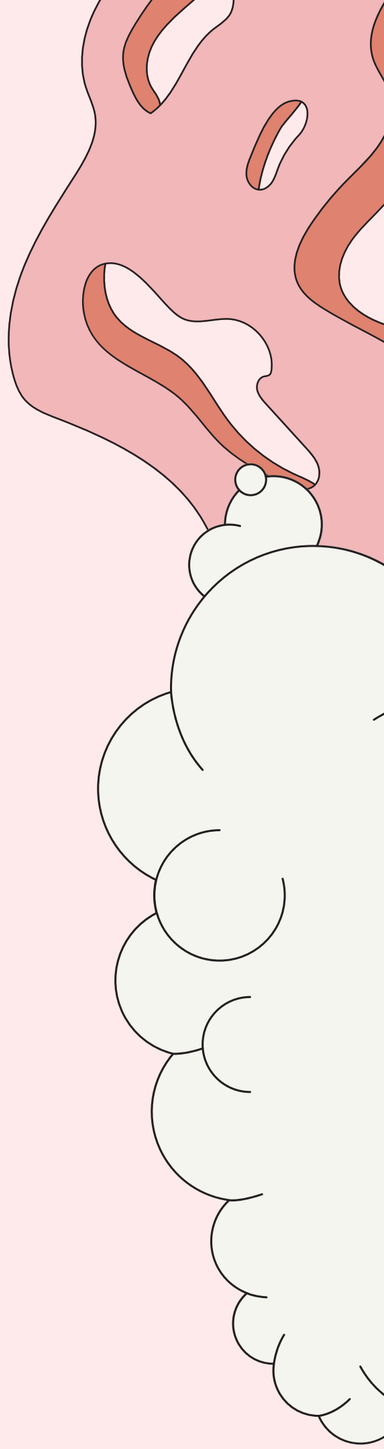


Cardiotocograms

- **Difficult to interpret**
- Interpreted Fetal Health Specialist

How does a data scientist play a role in reducing infant mortality?

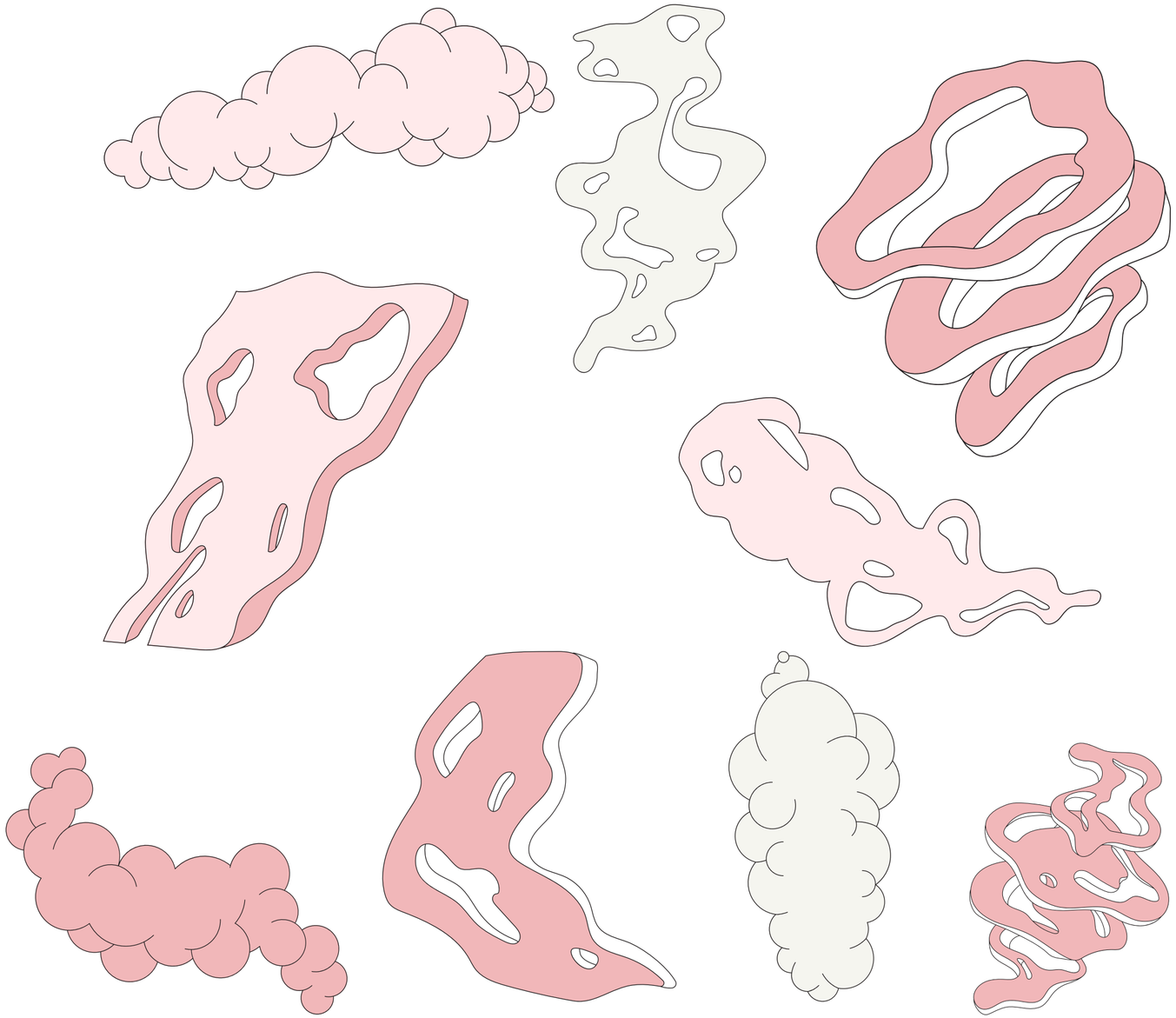
Building **machine learning predictions** about the **classification of fetal health status** using cardiotocography data



2

Data Understanding & Exploratory Data Analysis

Understand and take insight from data



DATASET

Contains

2126 rows & **22** Columns

link dataset:

https://www.kaggle.com/datasets/andrewmvd/fetal-health-classification

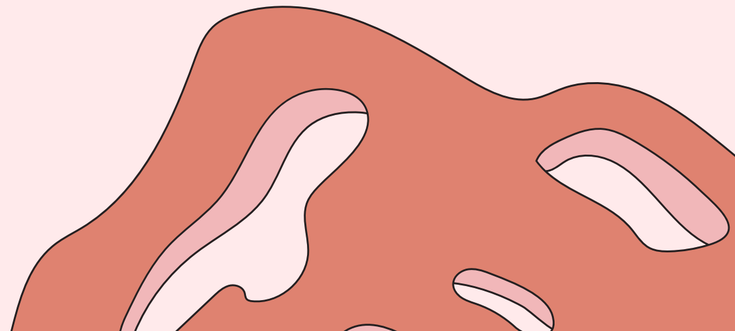
10 columns describe the **shape of the histogram**

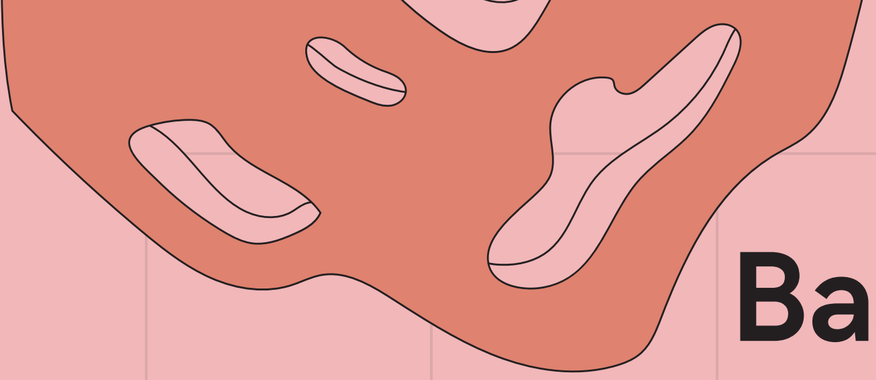
11 columns explain the **basic things in evaluating CTG** results

Fetal_health column as target
1 (**Normal**), 2 (**Suspect**) and 3 (**Pathological**)

Basic CTG Evaluation

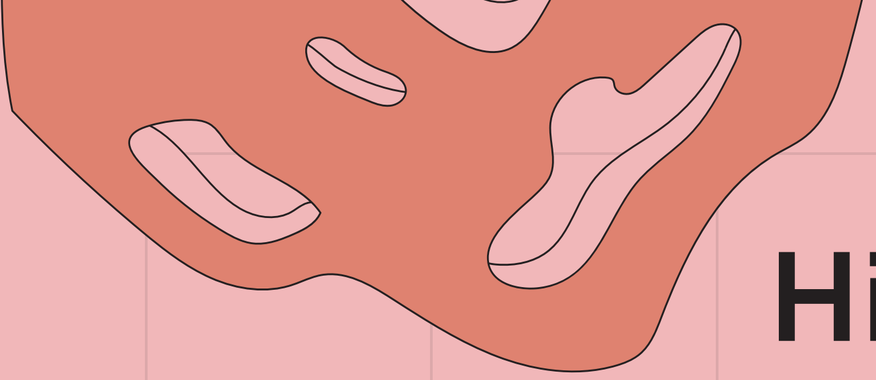
1. **Fetal Heart Rate**
2. **Uterus contraction**
3. **Acceleration**
4. **Deceleration**
5. **Variability**





Basic CTG Evaluation Columns

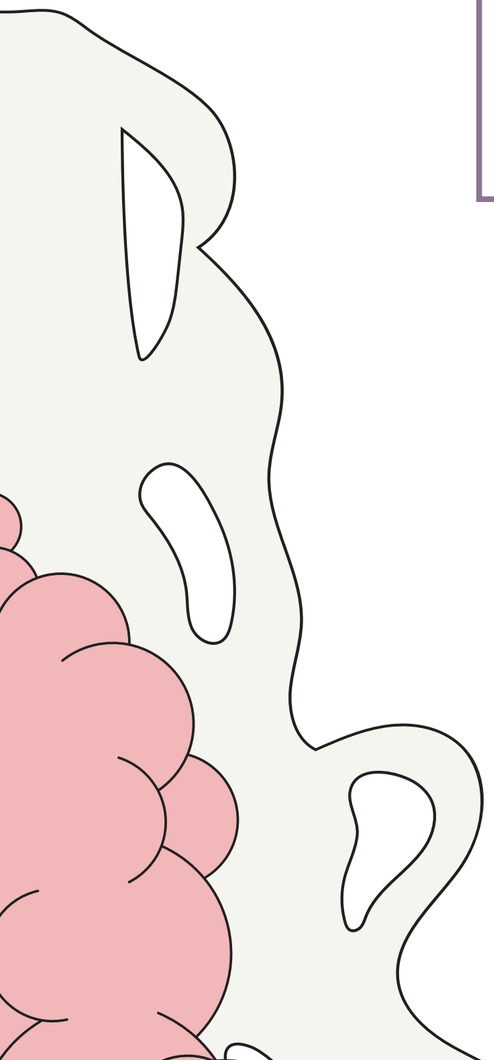
- **baseline_value**: FHR baseline (beats per minute)
- **accelerations**: Number of accelerations per second
- **fetal_movement**: Number of fetal movements per second
- **uterine_contractions**: Number of uterine contractions per second
- **light_decelerations**: Number of light decelerations per second
- **severe_decelerations**: Number of severe decelerations per second
- **prolongued_decelerations**: Number of prolonged decelerations per second
- **abnormal_short_term_variability**: Percentage of time with abnormal short term variability
- **mean_value_of_short_term_variability**: Mean value of short term variability
- **abnormal_long_term_variability**: Percentage of time with abnormal long term variability
- **mean_value_of_long_term_variability**: Mean value of long term variability



Histogram Shape Columns

- **histogram_width**: Width of FHR histogram
- **histogram_min**: Minimum (low frequency) of FHR histogram
- **histogram_max**: Maximum (high frequency) of FHR histogram
- **histogram_number_of_peaks**: Number of histogram peaks
- **histogram_number_of_zeroes**: Number of histogram zeros
- **histogram_mode**: Histogram mode
- **histogram_mean**: Histogram mean
- **histogram_median**: Histogram median
- **histogram_variance**: Histogram variance
- **histogram_tendency**: Histogram tendency

General Information & Data Cleaning



Duplicate	13

Missing values	0

Drop duplicates

Contains **2113** rows & **22** Columns

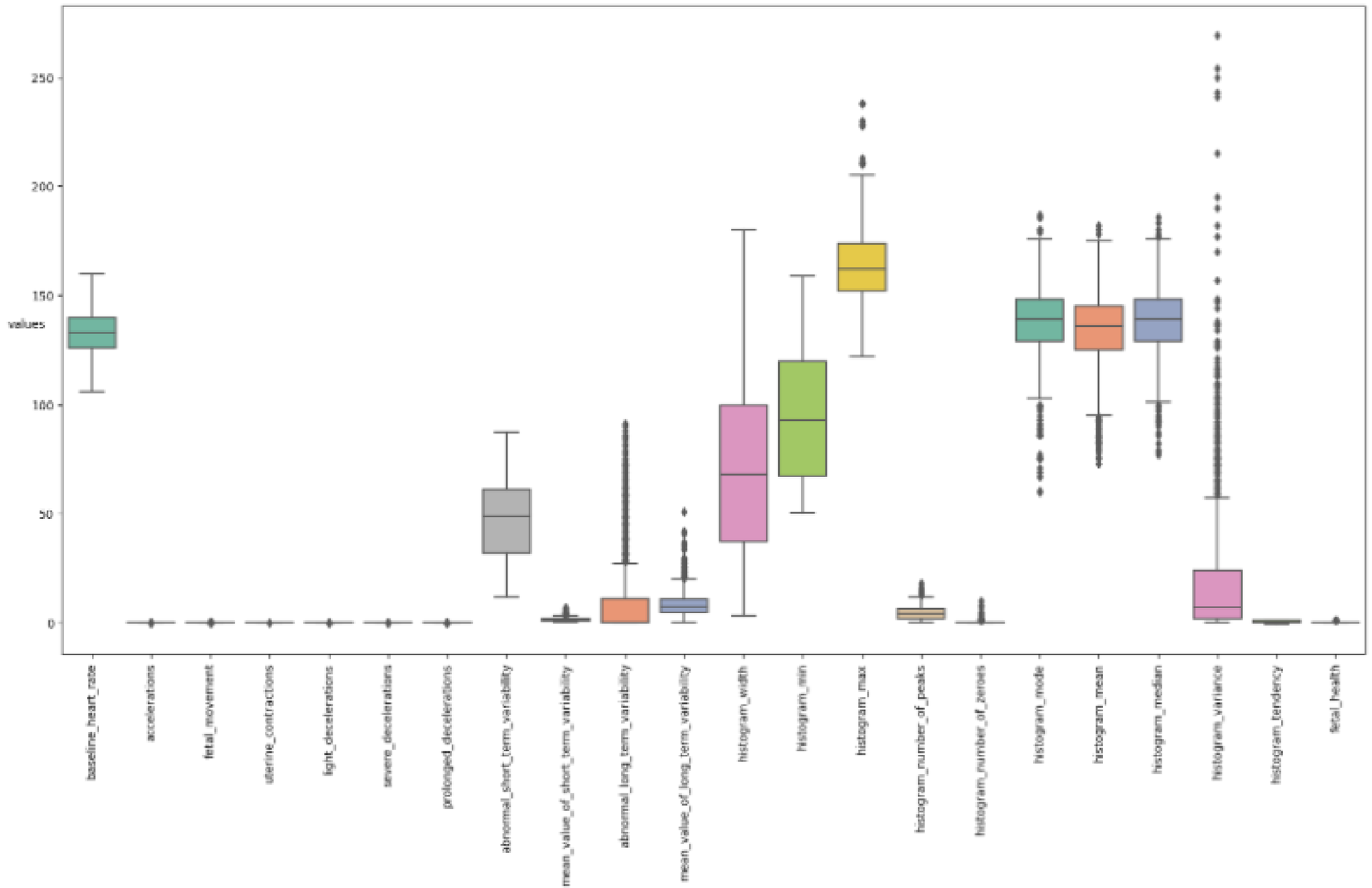
All **numerical** data type (float)

	count	mean	std	min	25%	50%	75%	max
baseline_heart_rate	2113.0	133.304780	9.837451	106.0	126.000	133.000	140.000	160.000
accelerations	2113.0	0.003188	0.003871	0.0	0.000	0.002	0.006	0.019
fetal_movement	2113.0	0.009517	0.046804	0.0	0.000	0.000	0.003	0.481
uterine_contractions	2113.0	0.004387	0.002941	0.0	0.002	0.005	0.007	0.015
light_decelerations	2113.0	0.001901	0.002966	0.0	0.000	0.000	0.003	0.015
severe_decelerations	2113.0	0.000003	0.000057	0.0	0.000	0.000	0.000	0.001
prolonged_decelerations	2113.0	0.000159	0.000592	0.0	0.000	0.000	0.000	0.005
abnormal_short_term_variability	2113.0	46.993848	17.177782	12.0	32.000	49.000	61.000	87.000
mean_value_of_short_term_variability	2113.0	1.335021	0.884368	0.2	0.700	1.200	1.700	7.000
abnormal_long_term_variability	2113.0	9.795078	18.337073	0.0	0.000	0.000	11.000	91.000
mean_value_of_long_term_variability	2113.0	8.166635	5.632912	0.0	4.600	7.400	10.800	50.700
histogram_width	2113.0	70.535258	39.007706	3.0	37.000	68.000	100.000	180.000
histogram_min	2113.0	93.564600	29.562269	50.0	67.000	93.000	120.000	159.000
histogram_max	2113.0	164.099858	17.945175	122.0	152.000	162.000	174.000	238.000
histogram_number_of_peaks	2113.0	4.077142	2.951664	0.0	2.000	4.000	6.000	18.000
histogram_number_of_zeroes	2113.0	0.325603	0.707771	0.0	0.000	0.000	0.000	10.000
histogram_mode	2113.0	137.454330	16.402026	60.0	129.000	139.000	148.000	187.000
histogram_mean	2113.0	134.599621	15.610422	73.0	125.000	136.000	145.000	182.000
histogram_median	2113.0	138.089446	14.478957	77.0	129.000	139.000	148.000	186.000
histogram_variance	2113.0	18.907241	29.038766	0.0	2.000	7.000	24.000	269.000
histogram_tendency	2113.0	0.318504	0.611075	-1.0	0.000	0.000	1.000	1.000
fetal_health	2113.0	1.303833	0.614279	1.0	1.000	1.000	1.000	3.000

Statistical Summary

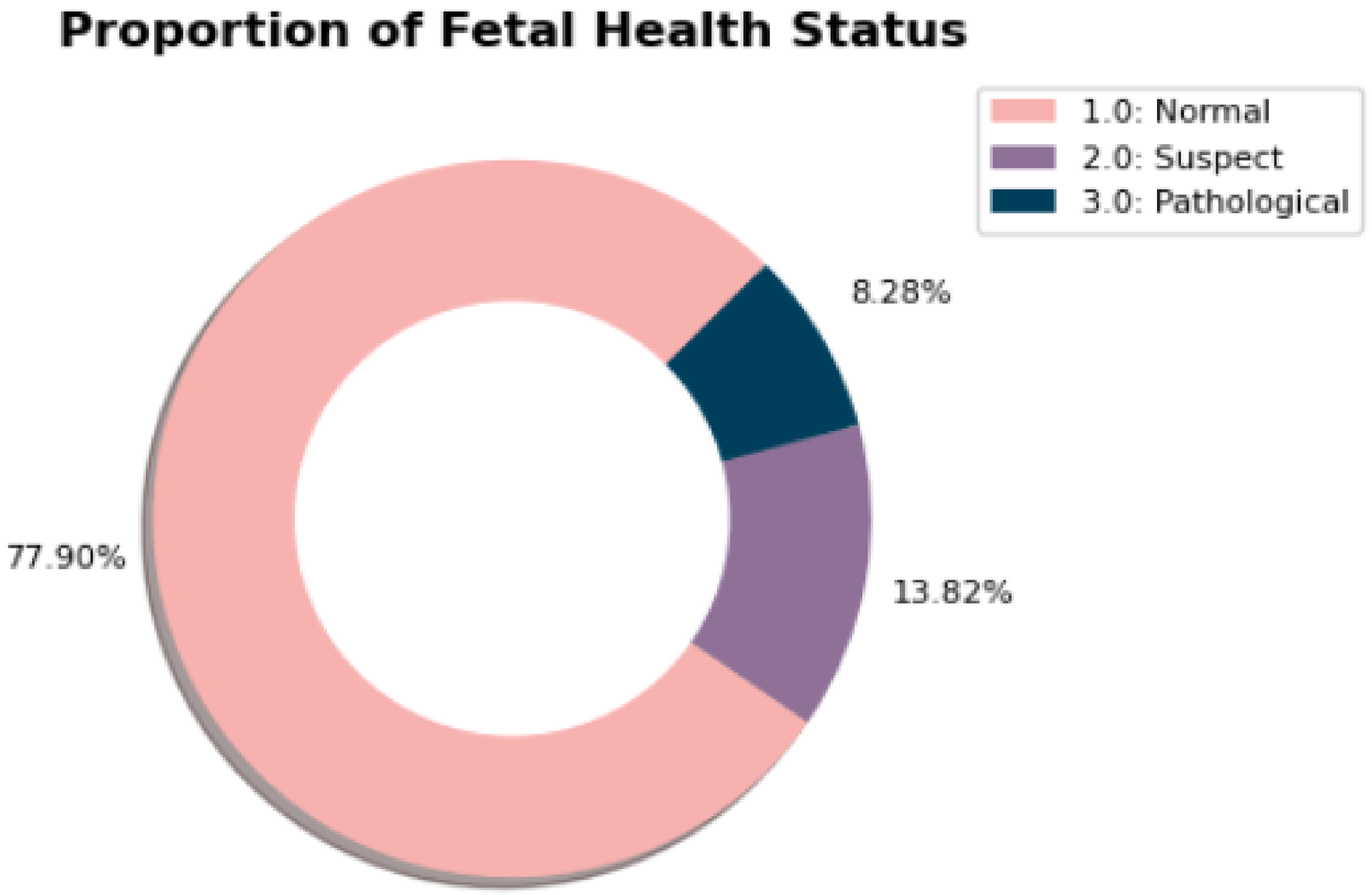
- Min-max values are mostly **making sense**
- Several columns are somewhat **symmetrical**, mean ~ median.

Box Plot Features



- There are **many columns with outliers**
- **The experts** marking its classification. So, **no need to droped**
- All the features are in **different ranges**

Target Column

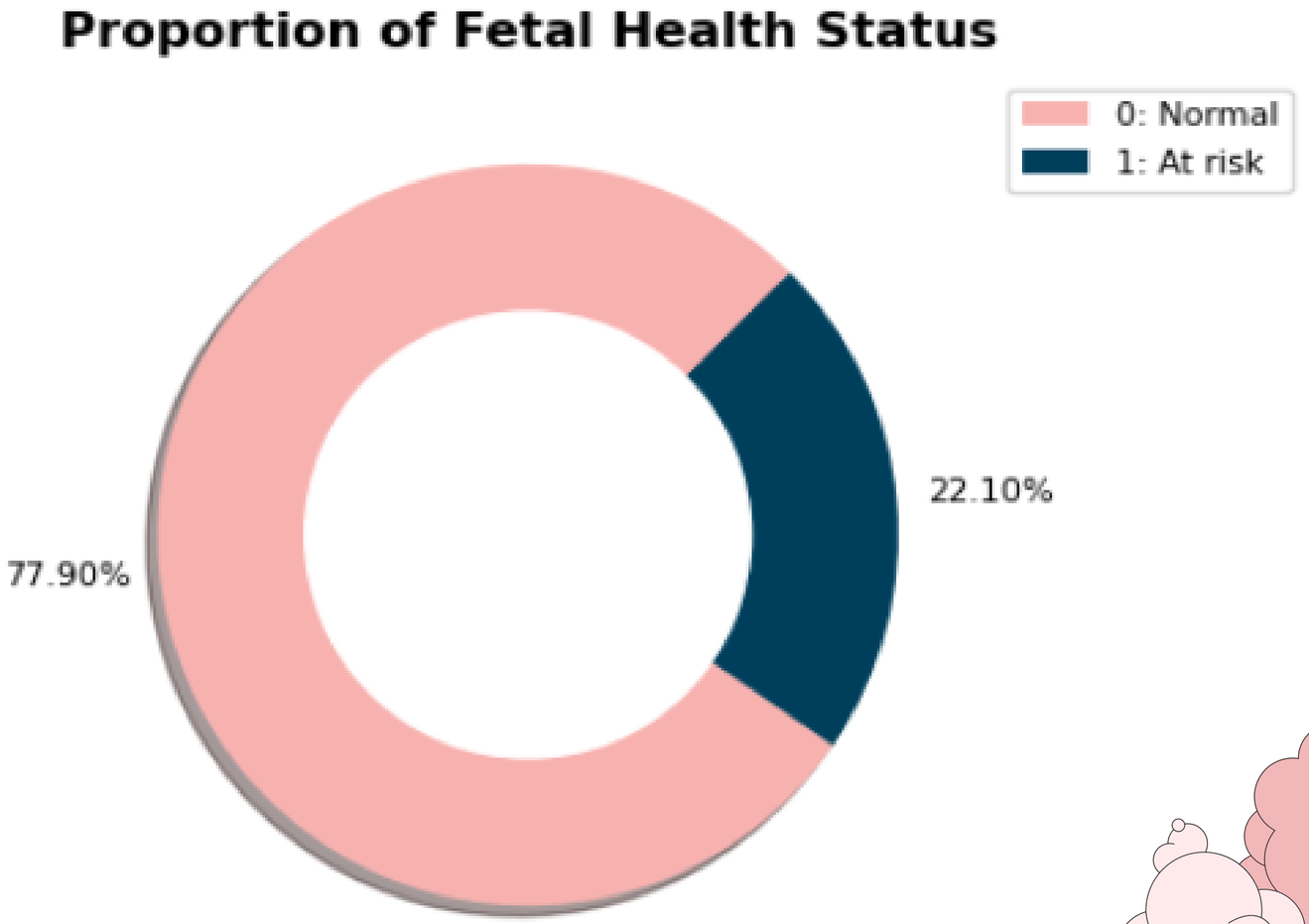


- **The majority of fetuses** in this dataset are **normal**
- **Normal** fetus as much as **1646**
- **292 suspected** fetuses
- **Pathological** fetuses as many as **175**

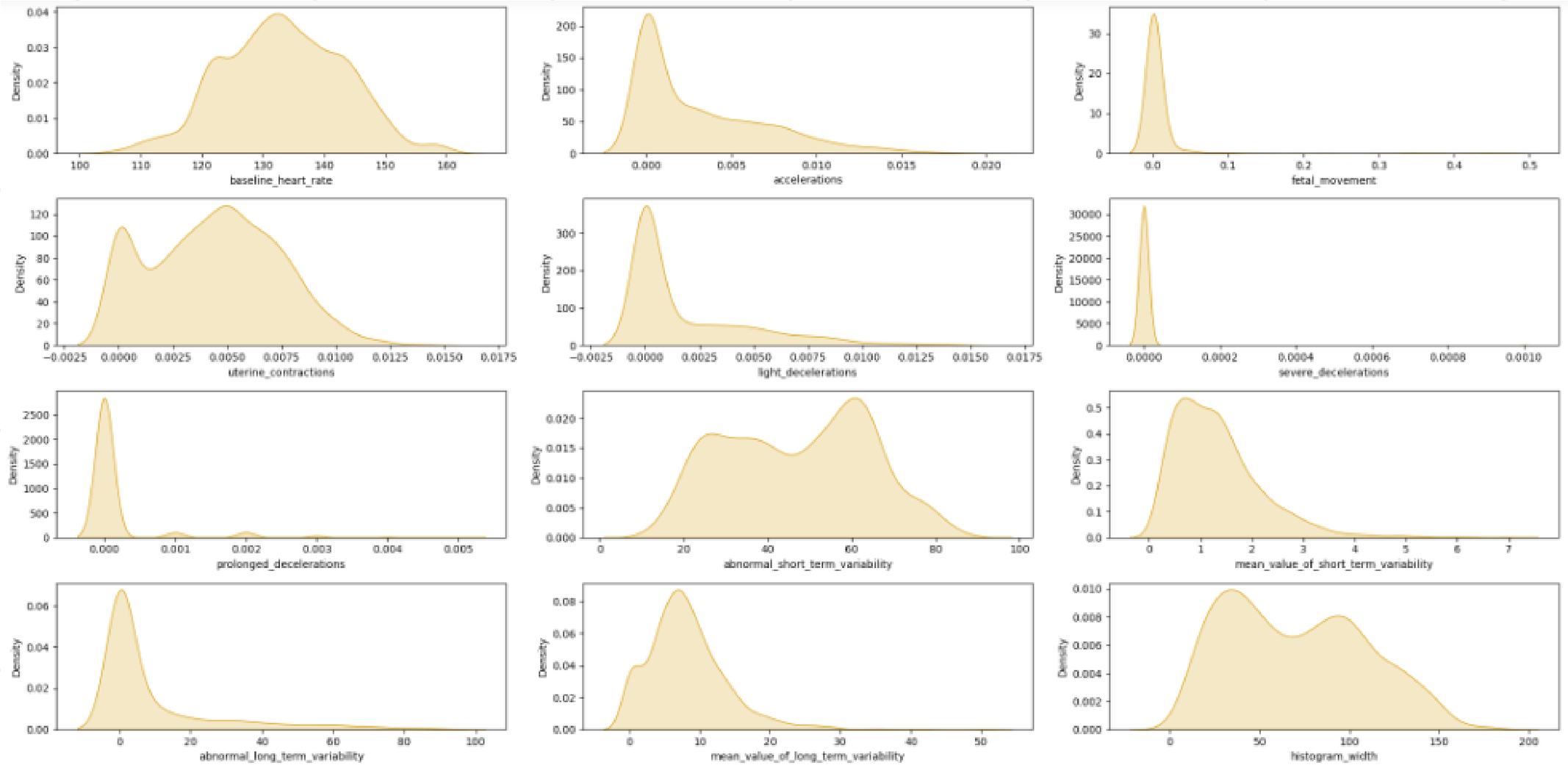
Feature Engineering Target Column

- The target column is **changed to binary**, by combining suspect and pathalogic

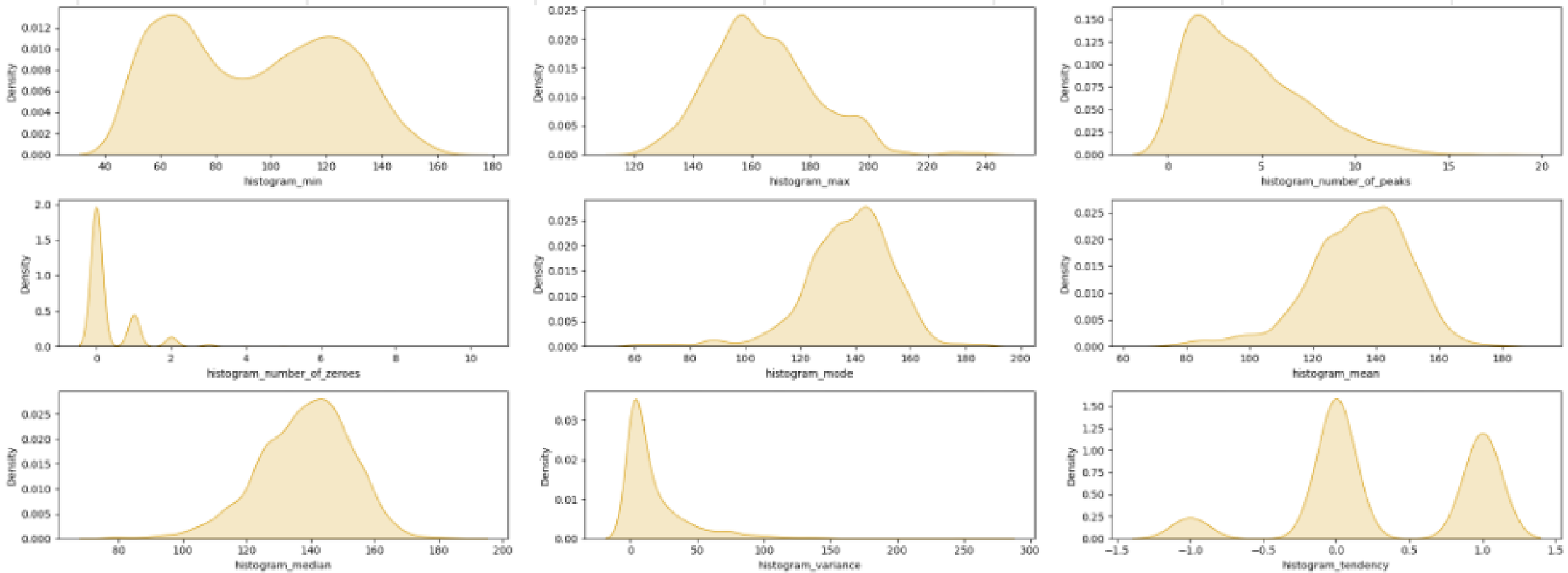
- A value of **0** means the fetus is **normal** (**1646**)
- A value of **1** means the fetus is **at risk (467)**



KDE Plot for Each Feature



KDE Plot for Each Feature



- Only **baseline_heart_rate**, **histogram_max**, **histogram_mode**, **histogram_mean**, **histogram_median** are close to **symmetrical**
- The **uterine_contractions**, **abnormal_short_term_variability**, **histogram_min** columns have a **bimodal distribution,** which has two peaks.

Observation KDE Plot

The trend of the fetus on this data has the following CTG results

- **Heart_rate**: 135 bpm
- **Accelerations**: 0
- **fetal_movement**: 0
- **uterine_contractions**: 0.005
- **light_decelerations**: 0
- **severe_decelerations**: 0
- **prolonged_decelerations**: 0
- **abnormal_short_term_variability**: 62 %
- **mean_value_of_short_term_variability**: 0.7
- **abnormal_long_term_variability**: 0 %
- **mean_value_of_long_term_variability**: 7

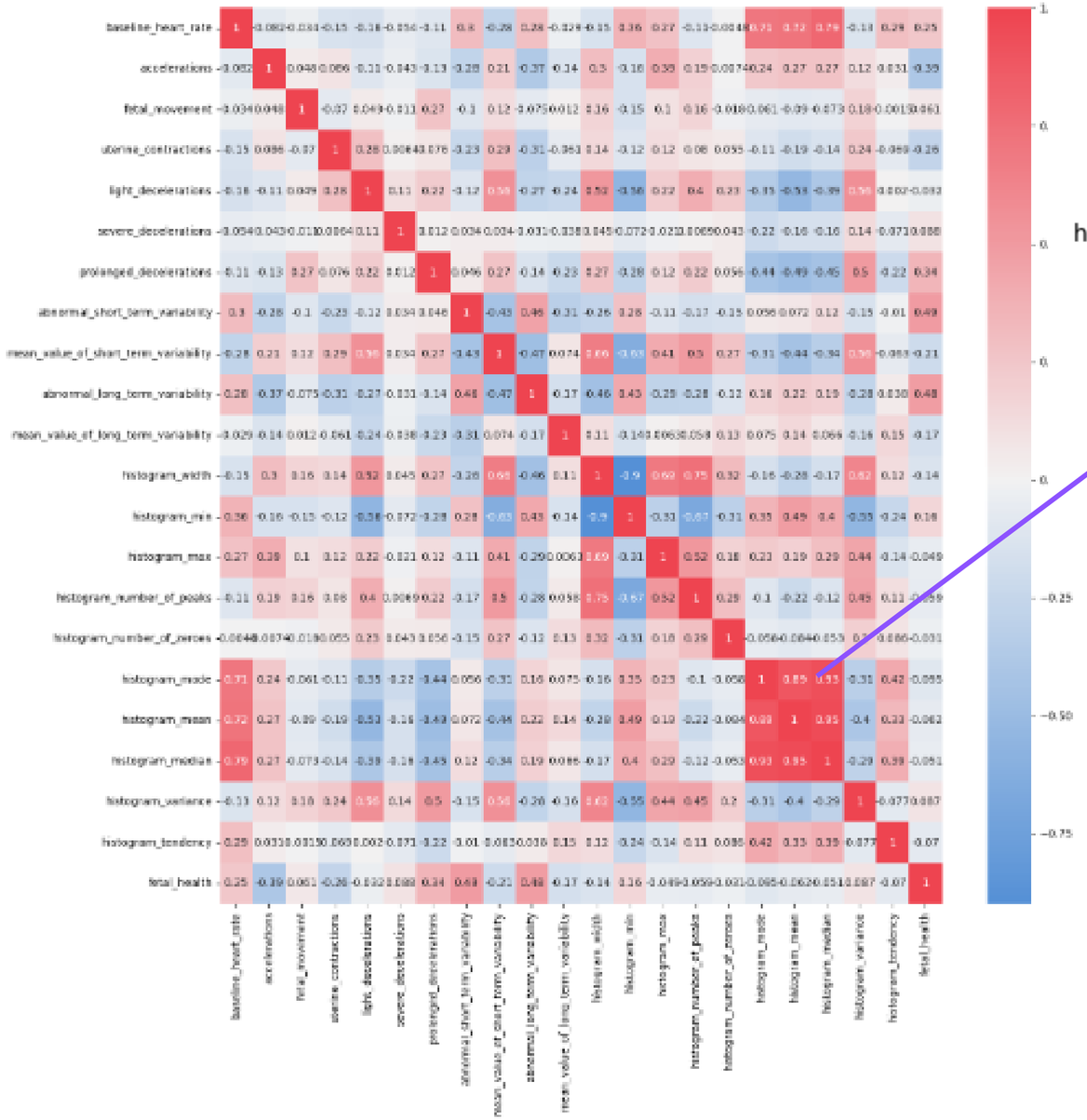
The majority of fetal health status in this dataset is **normal** so the **abnormal_short_term_variability** value should be **low** **and vice viersa.** This is according to AAFP.org.

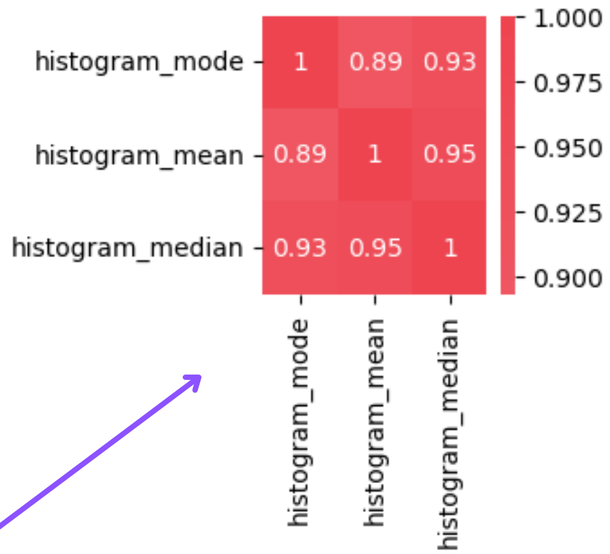
Heatmap correlation

Correlation with fetal_health

- **abnormal_short_term_variability** and **abnormal_long_term_variability** columns have a **strong positive** relationship
- The **baseline_heart_rate** and **prolonged_decelerations** columns have a **moderate positive**
- The **accelerations** and **uterine_contractions** columns have a **moderate negative**

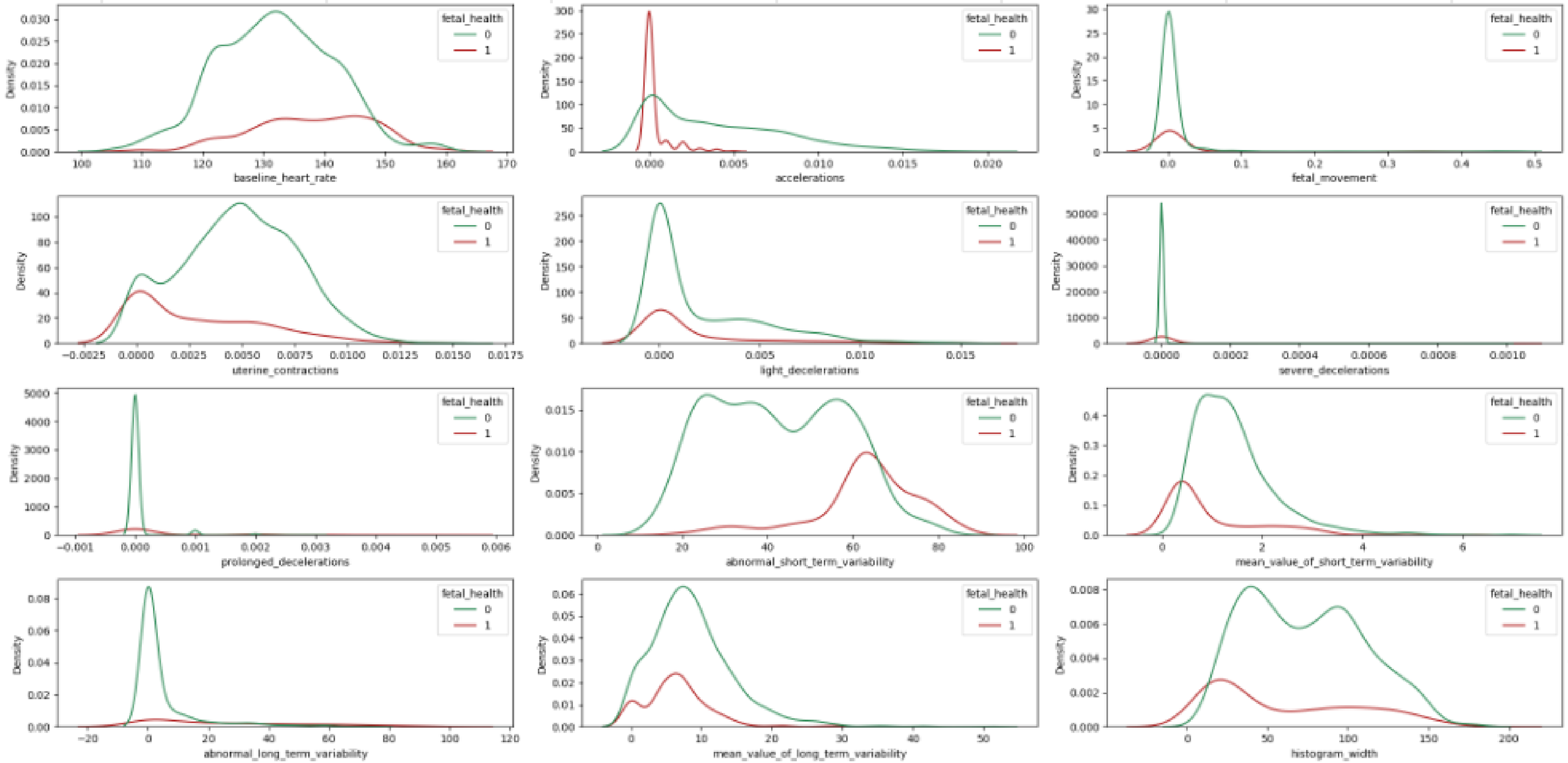
Heatmap correlation



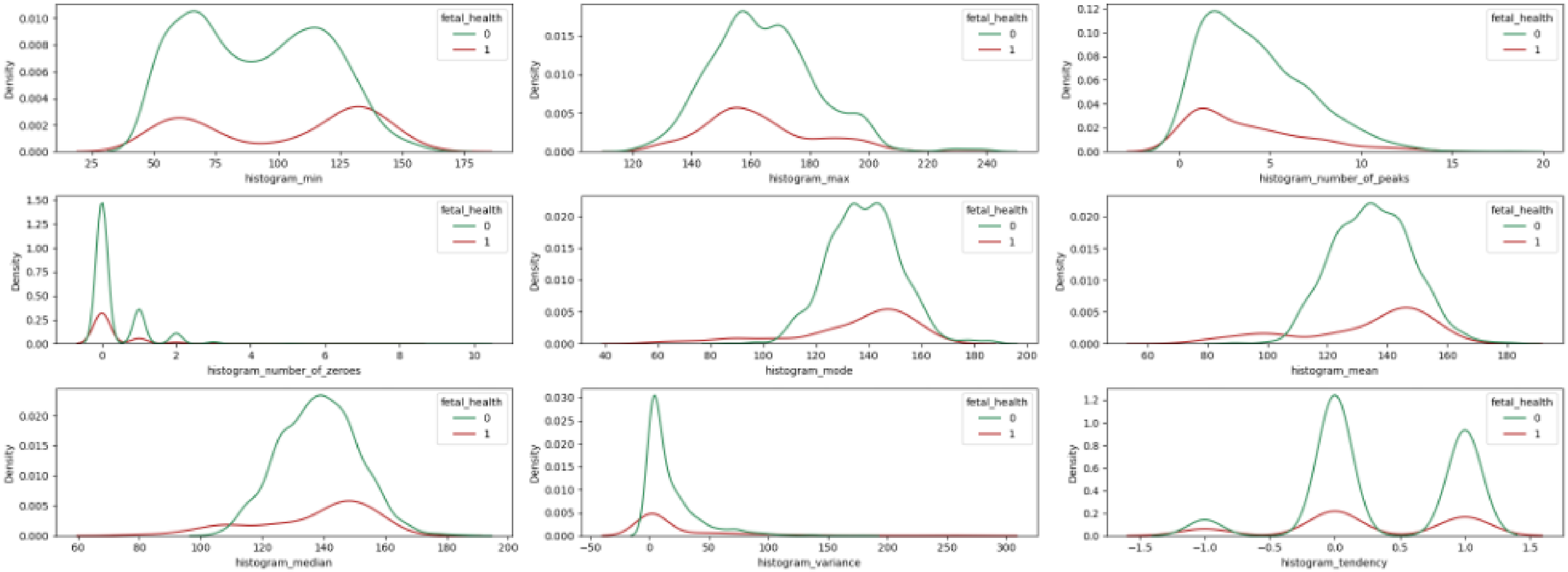


histogram_mode, **histogram_mean**, **histogram_median** are **very strongly positively** correlated, this shows that there is **redundancy** of information in the three columns

KDE plot based on Fetal Health

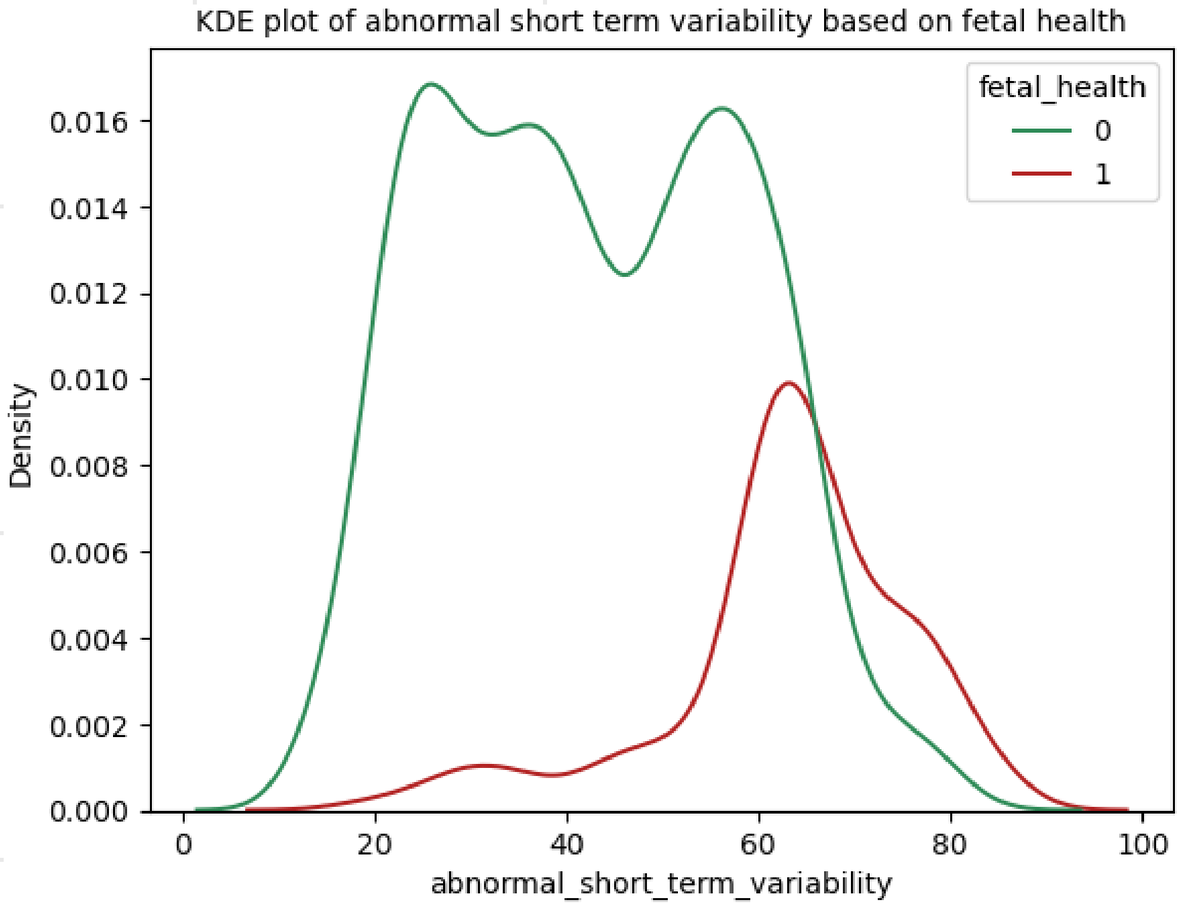


KDE plot based on Fetal Health



- The **baseline_heart_rate & prolonged deceleration** of a **healthy fetus** is **lower** than **risk fetuses**
- **uterine_contractions & acceleration** of **healthy fetuses** have a **higher** value than **risky fetuses**

KDE Plot of Abnormal Short Term Variability



- The **abnormal_short_term_variability** value of a **normal fetus** is **lower** than that of a **fetus at risk**
- **Normal fetuses** are most spread at +/- 23%, while **fetuses at risk** are +/- 65%.
- The overall **high abnormal short term variability** value (+/- 62%) **is caused** by **fetus at risk**.

CTG results are determined by signals from the **fetal heart rate** and **uterine contractions** (Rahmayanti et al. 2022).

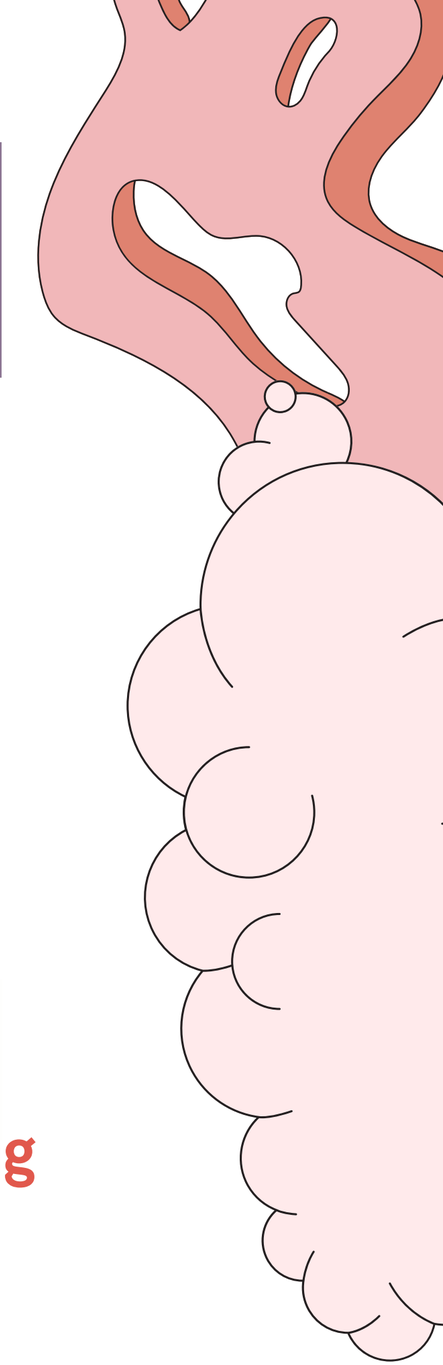
I want to prove this statement applies to this dataset !!

based on mean

Fetal Health	Fetal Heart Rate	Uterine Contraction
Normal	131 bpm	0,005
At risk	137 bpm	0,003

Is the difference **statistically significant**?

Let’s doing Hypothesis testing



Hypothesis Testing

Method: T-test **Alpha: 0.01**

H0

There is **no difference** in the average **fetal heart rate/uterine contractions** values for healthy and at-risk fetuses

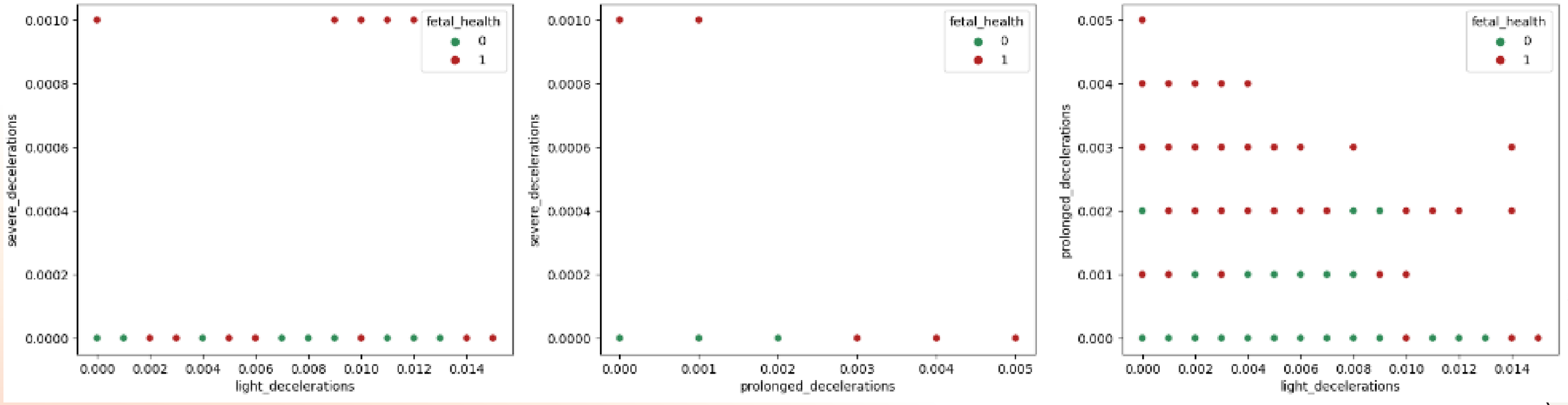
H1

There is **a difference** in the average **fetal heart rate/uterine contractions** values for healthy and at-risk fetuses

	p-value	alpha
Fetal Heart Rate	3×10^{-31}	0.01
Uterine Contraction	5×10^{-35}	0.01

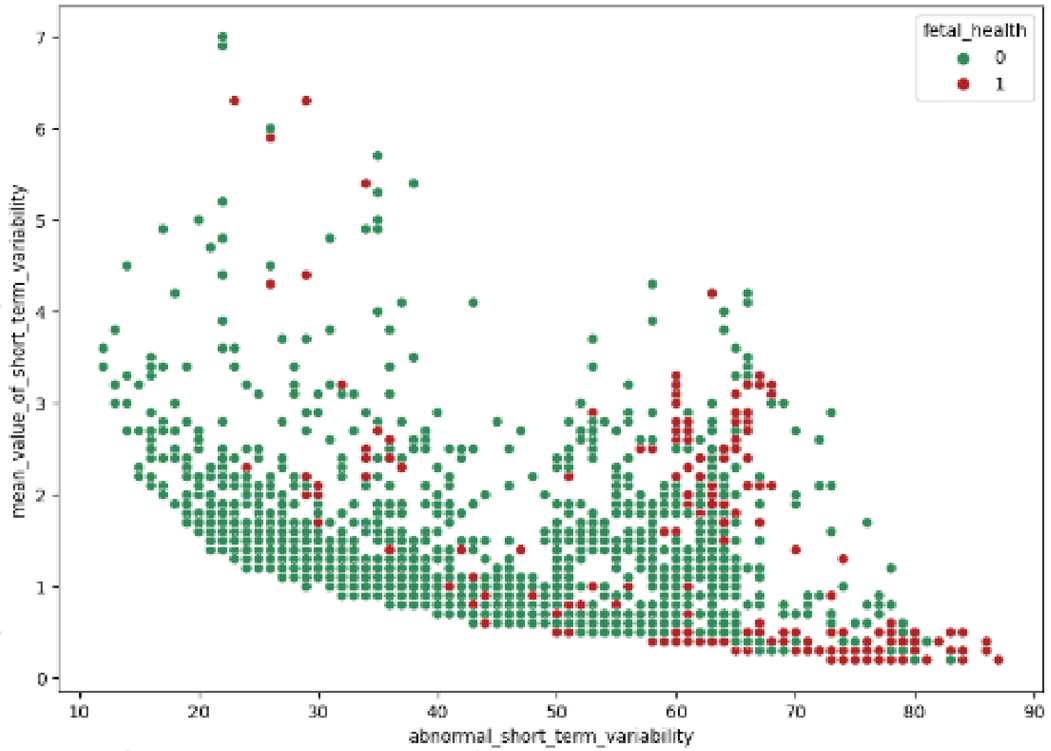
The ***p-value*** of **fetal heart rate** and **uterine contractions** is **below** 0.01, so **H0 is rejected** and **H1 is accepted**

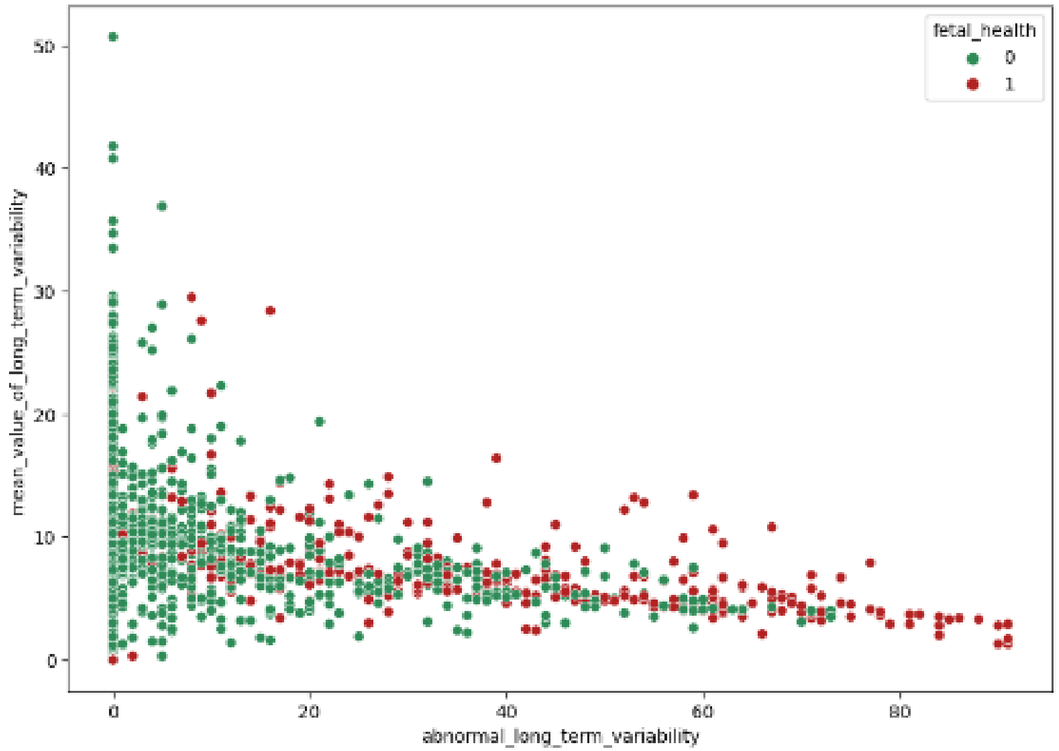
Relationship of deceleration columns



- The deceleration columns are of discrete numeric type
- With the same **light_decelerations** and **prolonged_decelerations** values, a **normal** fetus will have a **lower** **severe_decelerations** value than a **risky** fetus.
- With the same **light_decelerations** value a **normal** fetus will have a **lower** **prolonged_decelerations** value than a **risky** fetus and vice versa.

Relationship between variability columns



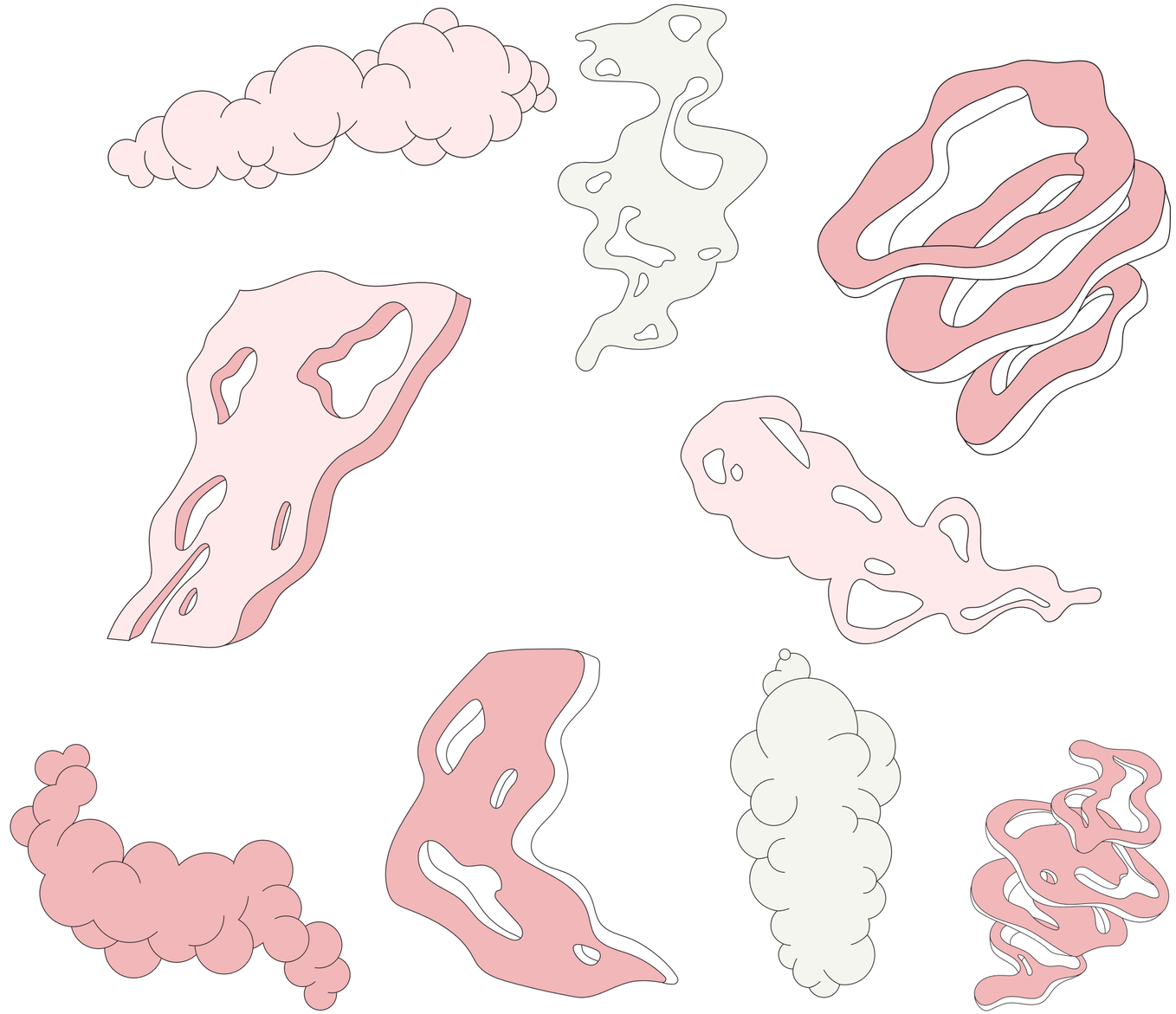


- With the same **mean_value_of_short_term_variability** value, **risky fetuses** tend to have **higher abnormal_short_term_variability** values than **healthy fetuses**, likewise for **long term** ones.
- The **mean value variability** column has **negative correlation** with the **abnormal variability column** for both **long term** and **short term**, but the correlation for **short term** is **clearer**

3

Modeling

Supervised machine learning classification



6 Classification Model

1. Logistic regression
2. K-Neareast neighbors
3. Decission Tree
4. Random Forest
5. Naive Bayes Classifier
6. Support Vector Machine

Target

0: Normal 1: At risk

Two types of **False:**

- **False Positive**: we predict that the fetus is at risk, but it is **actually normal**
- **False Negative**: we predict that is normal, but it's **actually a risk**

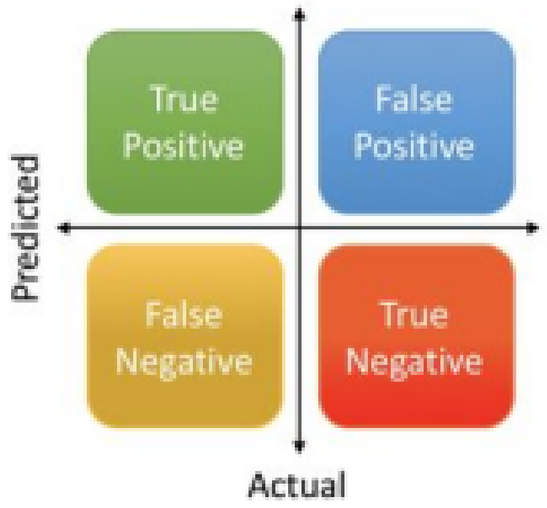
Evaluation matric in classification:

- Precision **- Recall** **- F1-Score** **- Accuracy**

$$\textbf{Precision} = \frac{\text{True Positive}}{\text{Actual Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$

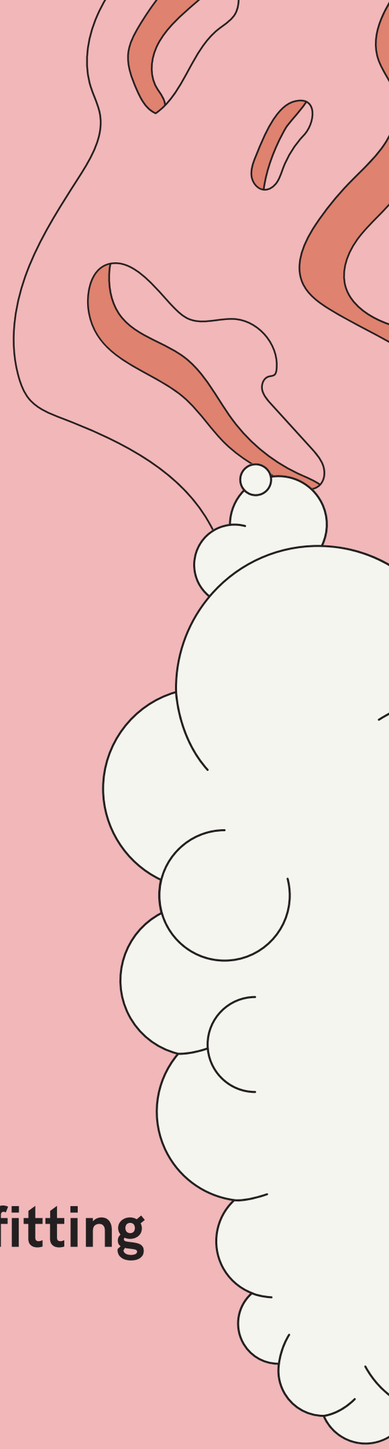
$$\textbf{Recall} = \frac{\text{True Positive}}{\text{Predicted Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

$$\textbf{Accuracy} = \frac{\text{True Positive + True Negative}}{\text{Total}}$$



Baseline Model Performance

Model	Recall Training	Recall Testing
Logistic regression	63.60%	67.14%
K-Neareast neighbors	77.98%	72.85%
Decission Tree	99.38%	80.71%
Random Forest	**99.60%**	**82.85%**
Naive Bayes Classifier	79.51%	77.85%
Support Vector Machine	52.90%	55.00%



The best baseline model is **Random forest**, but there are concerns about **overfitting**

Pre-processing steps

Multicolinearity Handling

DROP 2 COLUMNS

Keep the histogram_mode

01.

02.

Scaling

STANDARDSCALER

Split train, test. Transform it

03.

Imbalance Handling

SMOTE

Only Training data

04.

Hyperparameter Tuning

2 PARAMETERS

25 Combination

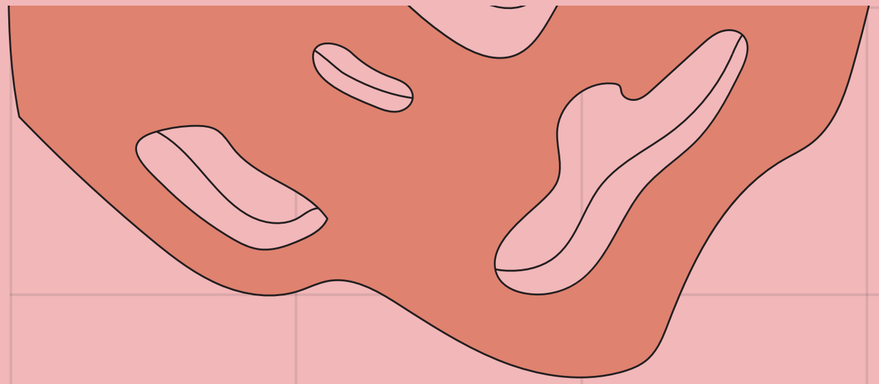
05.

The best performance model

Combination Pre-processing

1. **Multicolinearity Handling**
2. **Scaling (StandardScaler)**
3. **Imbalance Handling (SMOTE)**
4. **Hyperparameter Tuning**

Combination	Recall Training	Recall Testing
All pre-processing	99.91%	88.57%
Scaling + Imbalance + HP Tuning	99.91%	89.28%
Imbalance +HP Tuning	**99.91%**	**92.14%**
Imbalance	99.91%	91.42%
HP Tuning	99.69%	84.28%



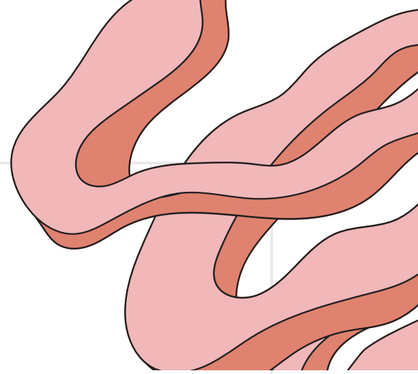
The Best Model

Random Forest

min_samples_split=4, n_estimators=200

Evaluation

0 : Normal **1 : At risk**



Model	Recall Training	Recall Testing	Accuracy Training	Accuracy Testing
Random Forest min_samples_split=4, n_estimators=200	99.91%	92.14%	99.86%	96.05%

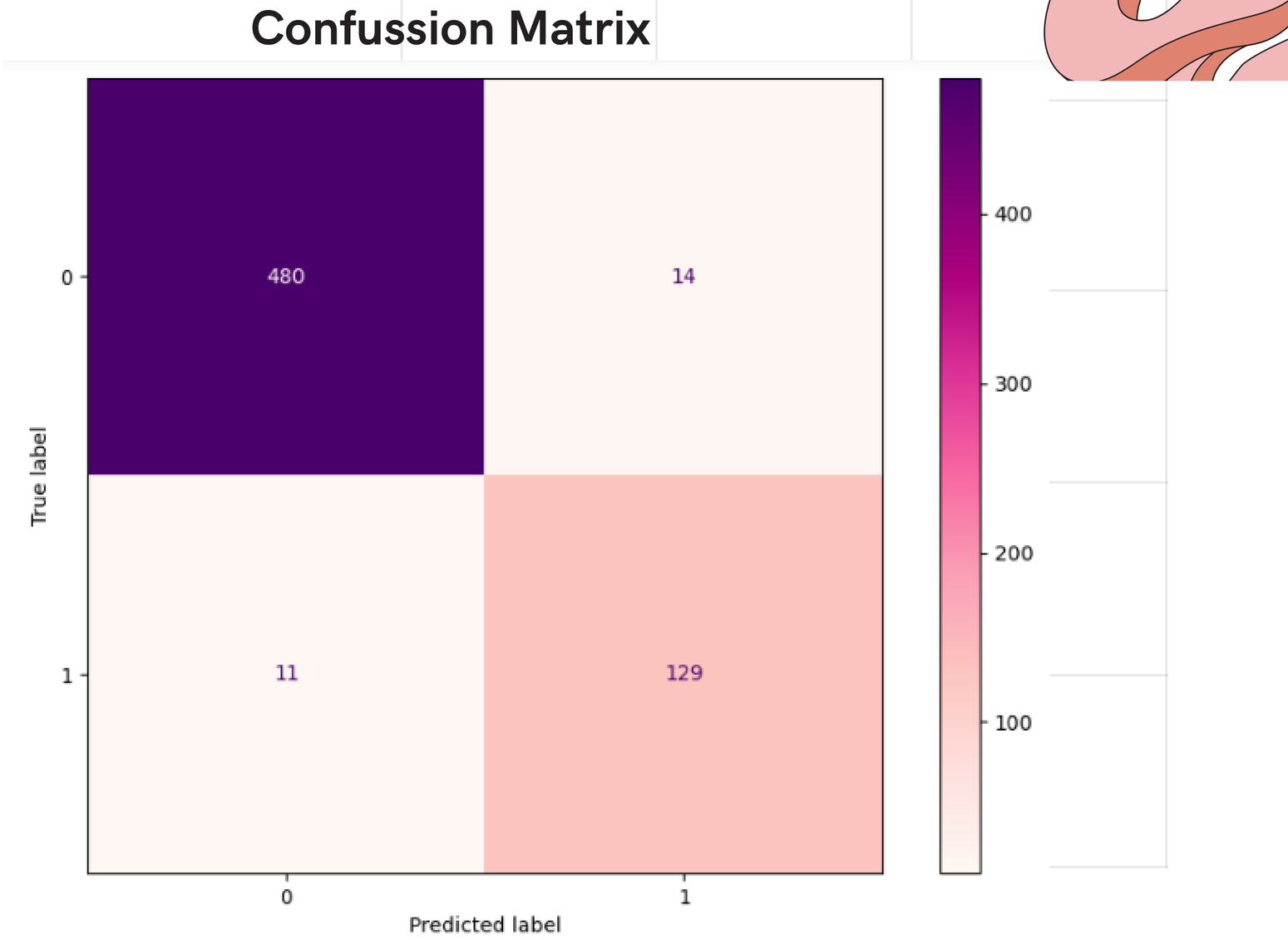
The model performs well in **training** and **testing** data, so it is **not overfitting**

- Recall= **92.14%** means that out of **100 fetuses** that are **truly at risk**, our model is **able to detect 92** of them.
- Accuracy= **96.05%** means that our model is able to **correctly predict** fetal health in **96 out of 100 fetuses**.

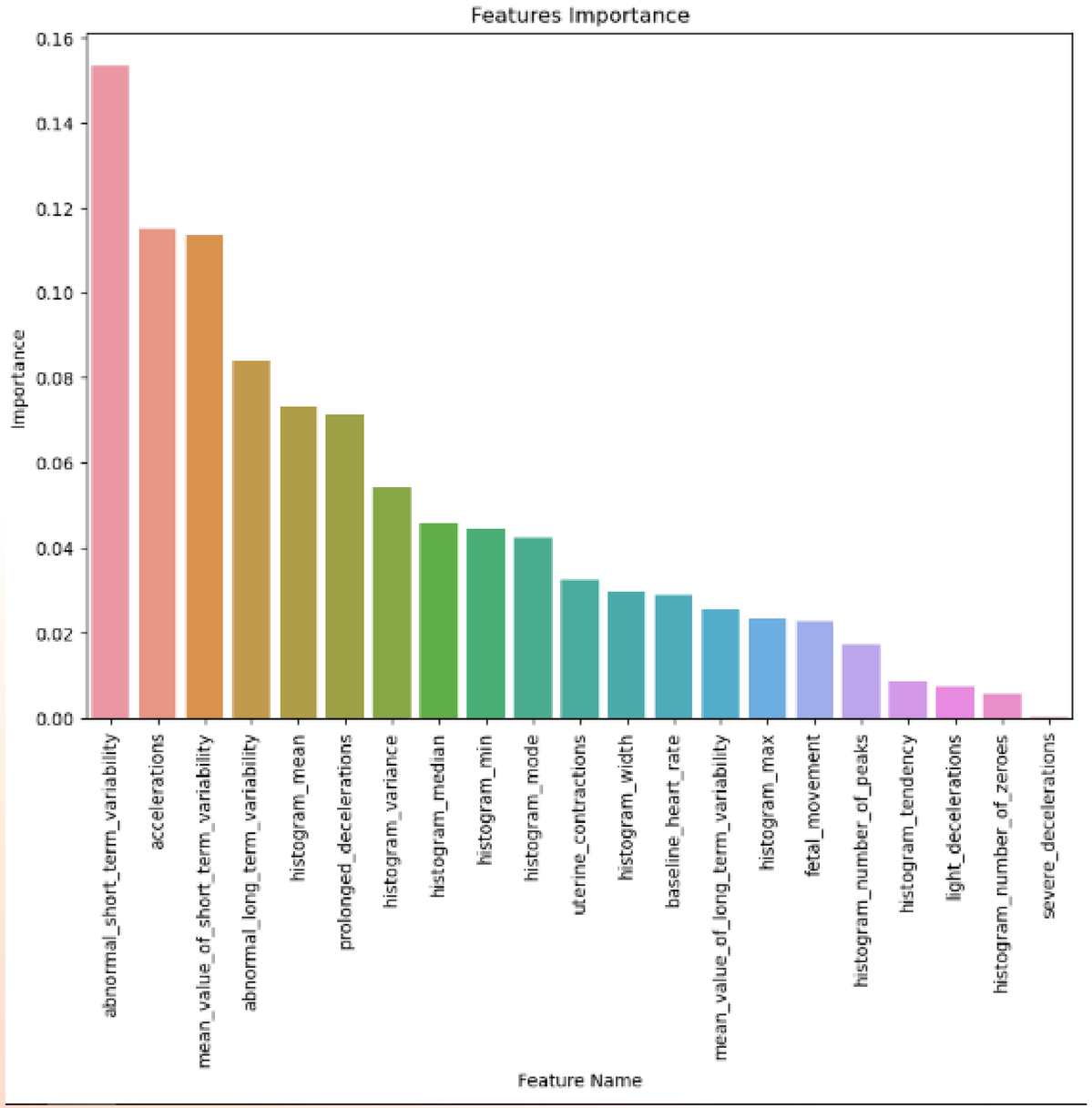
Evaluation

0 : Normal **1 : At risk**

- **True Negative:** 480
- **True Positive:** 129
- **False Positive:** 14
- **False Negative: 11**



Features Importance



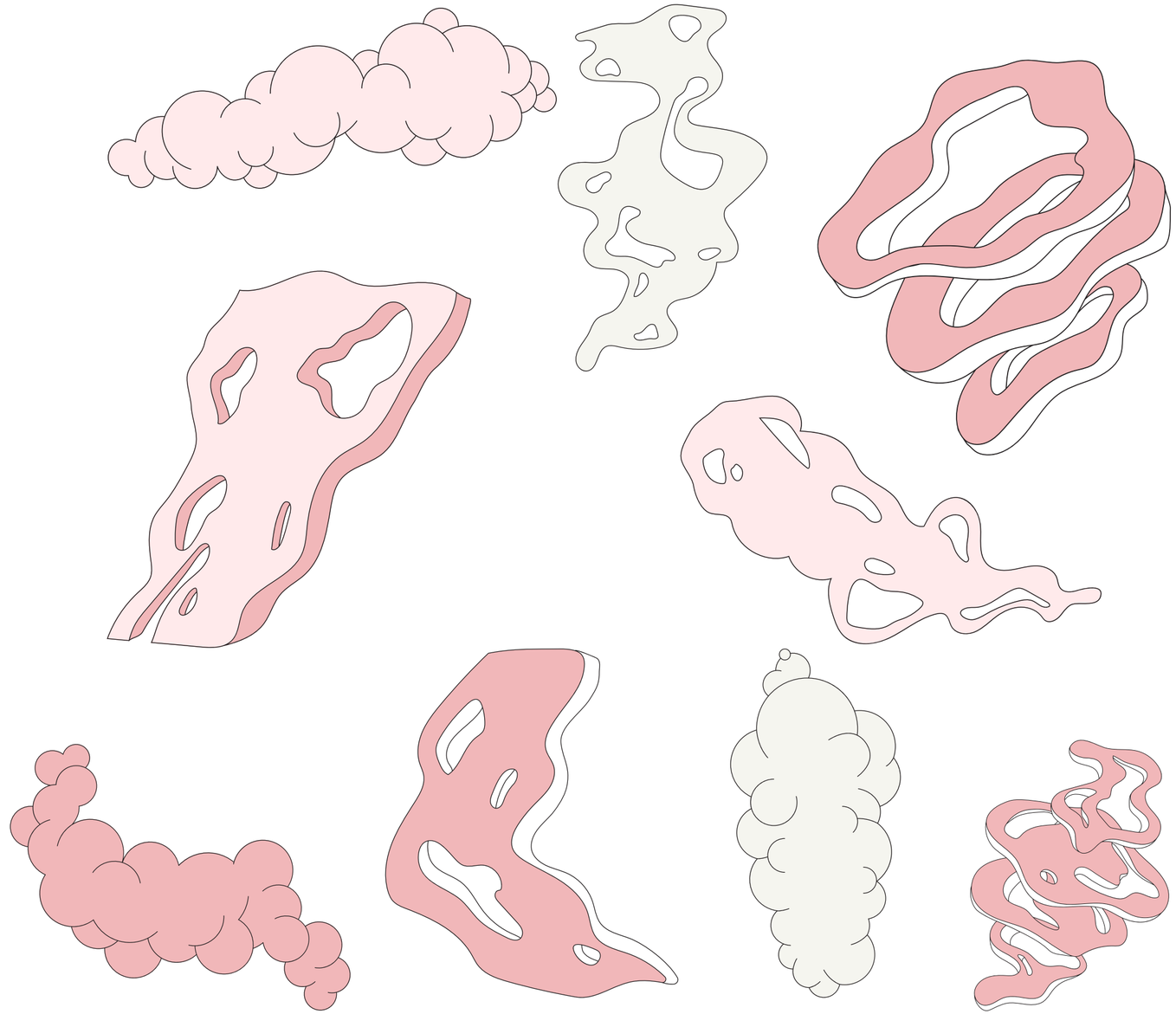
The **strongest predictor** of fetal health:

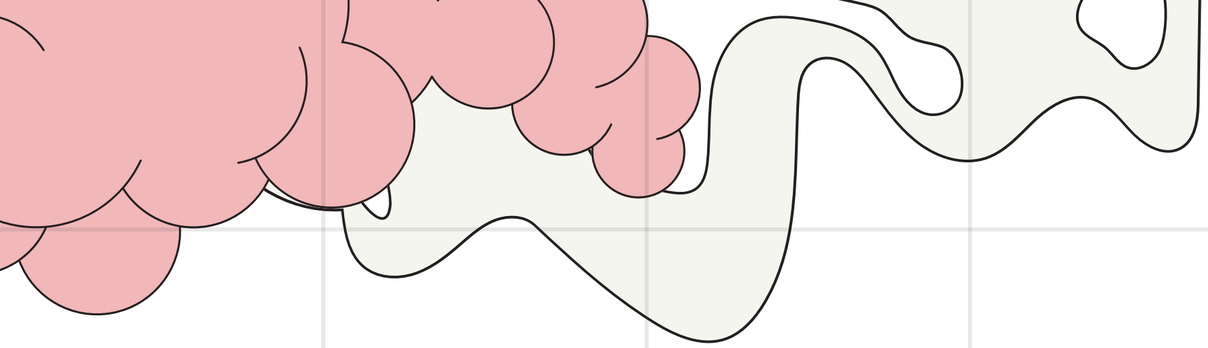
- abnormal_short_term_variability
- accelerations
- mean_value_of_short_term_variability

4

Recomendation

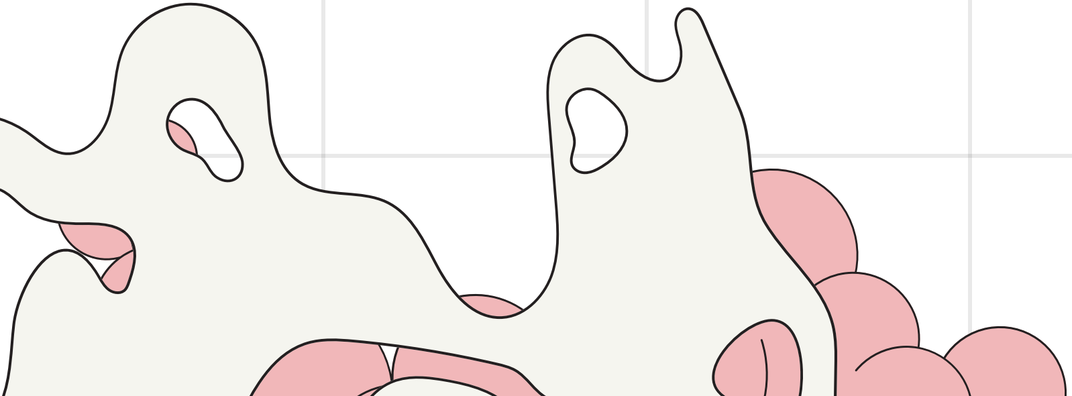
Conclusion and
Recomendation





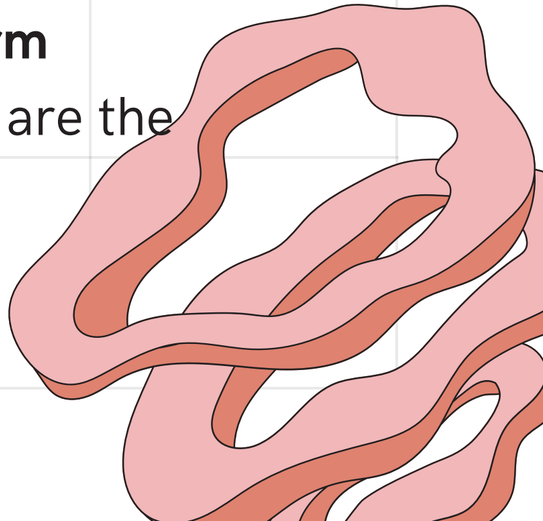
Conclusion

A **machine learning model can be built** to predict fetal health status using cardiotocography data with **good performance.** The **random forest** model with parameters **min_samples_split: 4, n_estimators: 200** is the best model with a **recall** value of **91.97%** and an **accuracy** of **94.79%.** It is hoped that the **automation of cardiotocogram analysis** using machine learning and the knowledge of obstetricians can **reduce fetal death rate** .



Recomendation

1. I suggest **applying machine learning in automating cardiotocogram analysis** so that it makes it **easier** for obstetricians and **can analyze it quickly** so that cardiotocography tests can be carried out **more frequently**.
2. It is necessary to **add a dataset** with **various types of cardiotocograph** equipment and **geographical locations,** this will provide a universal picture and more insight into the machine learning model so that its performance **will be better and more valid**.
3. I **recommend** that healthcare providers pay attention to **abnormal short term variability**, **acceleration**, and **mean value of short term variability** as these are the **strongest predictors** of fetal health risk.



Thank You.

Reach Me

https://www.linkedin.com/in/azri-ahza

https://github.com/Ahzaazr

azrhza9@gmail.com

http://wa.me/628136079741