STATEMENT OF WORK (V2)

MARCOS BITTENCOURT

DECEMBER 1, 2020

PARTH CHOTAI - 100802846

Table of Contents

1.1 Problem Definition

Reduction of child mortality is reflected in several of the United Nations' Sustainable Development Goals and is a key indicator of human progress. The UN expects that by 2030, countries end preventable deaths of new-borns and children under 5 years of age, with all countries aiming to reduce under-5 mortality to at least as low as 25 per 1,000 live births.

Deaths during and following pregnancy and childbirth is 295,000 (as of 2017). Most of these deaths (94%) occurred in low-resource settings, and most could have been prevented.

1.2 Objective / Rationale Statement

Cardiotocograms (CTGs) are a simple and cost accessible option to assess fetal health. This allows healthcare professionals to act in order to prevent child and maternal mortality.

We will predict **fetal_health** from CTGs data. The goal is to be able to respond to the risk of death in advance.

1.3 Dataset Information

The dataset (which is taken from Kaggle) contains 2126 records of features extracted from Cardiotocogram exams, the CTGs were also classified by three expert obstetricians and a **consensus classification label** assigned to each of them. Classification was both with respect to a **morphologic pattern (A, B, C. ...)** and to a **fetal class (N, S, P)**. Therefore, the dataset can be used either for 10-class or 3-class experiments.

Inputs

The dataset contains a total of 21 inputs described below:

1. baseline value
2. accelerations
3. fetal_movement
4. uterine_contractions
5. light_decelerations
6. severe_decelerations
7. prolongued_decelerations

8. abnormal_short_term_variability
9. mean_value_of_short_term_variability
10. percentage_of_time_with_abnormal_long_term_variability
11. mean_value_of_long_term_variability
12. histogram_width
13. histogram_min
14. histogram_max
15. histogram_number_of_peaks
16. histogram_number_of_zeroes
17. histogram_mode
18. histogram_mean
19. histogram_median
20. histogram_variance
21. histogram_tendency

Target Variable

It uses the **fetal_health** as the **target variable**. As mentioned above, fetal class is classified according to 3 situations (**N** — Normal, **S** — Suspect or **P** — Pathological).

1.4 Dataset Constraints

After initial EDA, it has been found that **target variable** distribution is imbalanced. In order to solve the problem related to the imbalance of the dataset, we will need to use sampling to equalize the number of samples for each of the classes **N**, **S** and **P**.

1.5 Testing Process

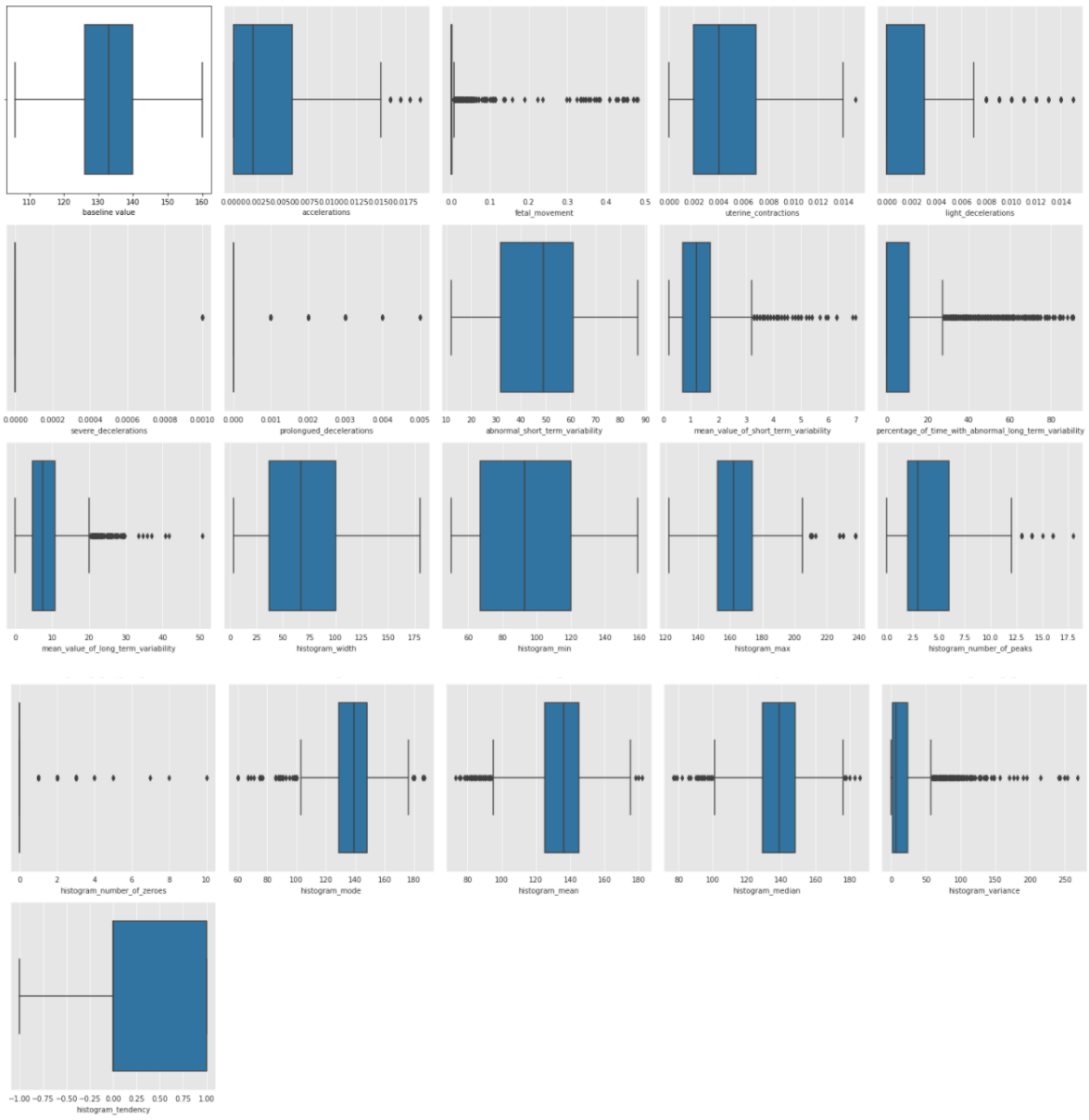
- Invariance Tests
- Directional Expectation Tests
- Minimum Functionality Tests (aka data unit tests)

1.6 Project Plan

The table below contains the project tasks and their estimated completion dates.

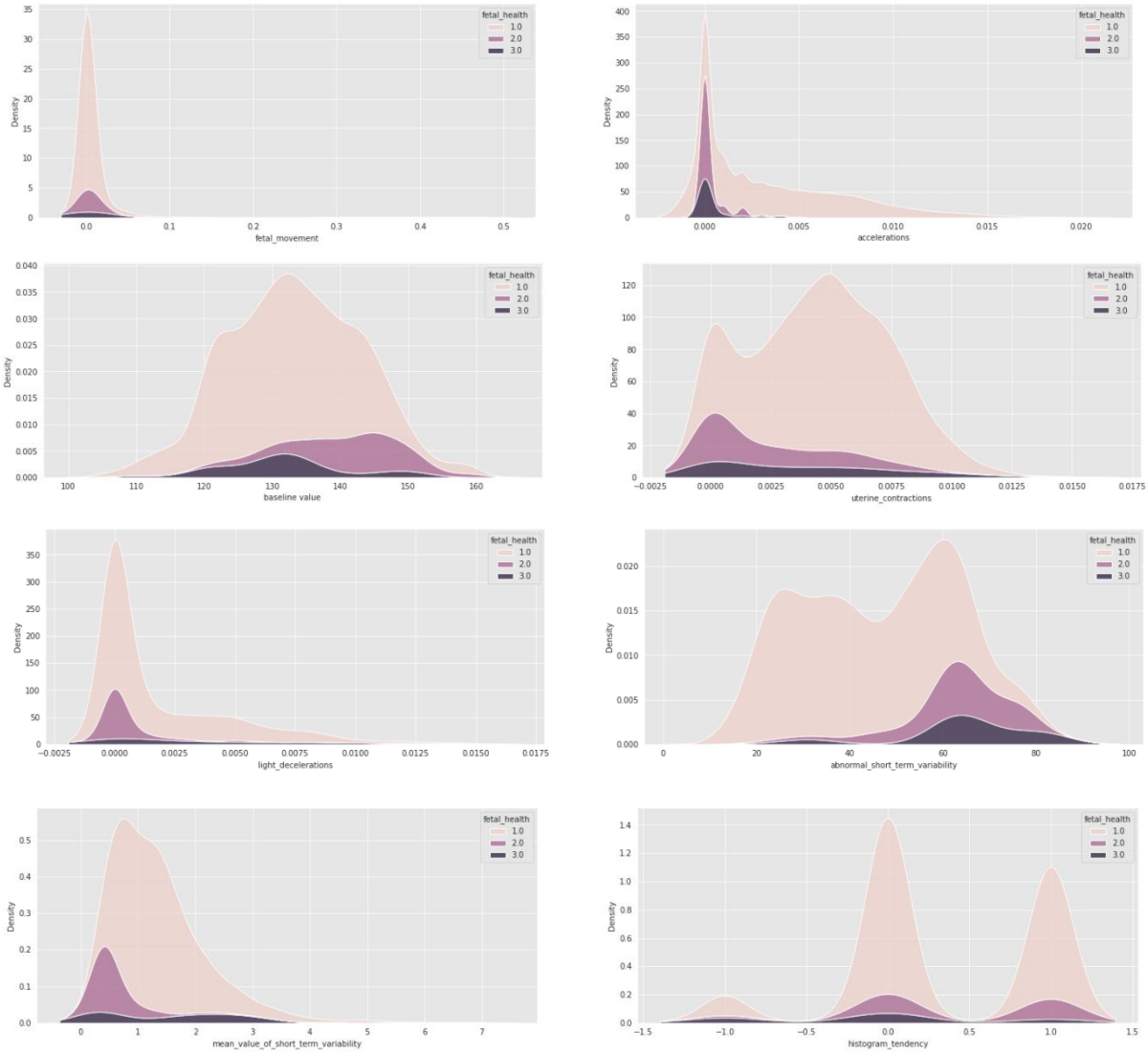
Tasks	Details	Delivery Date
Statement of Work (V1) **(Business Understanding & Problem Discovery)**	Problem Definition, Data Requirements and Assumptions	6th November, 2020
Statement of Work (V2) **(Data Acquisition & Understanding)**	Exploratory Data Analysis, Data Cleaning and Feature Engineering	23rd November, 2020
Modeling & Prototyping	Data Manipulation, Preliminary Model Building and Evaluation	23rd November, 2020
Deployment	Final Delivery	18th December, 2020

2.1 Exploratory Data Analysis



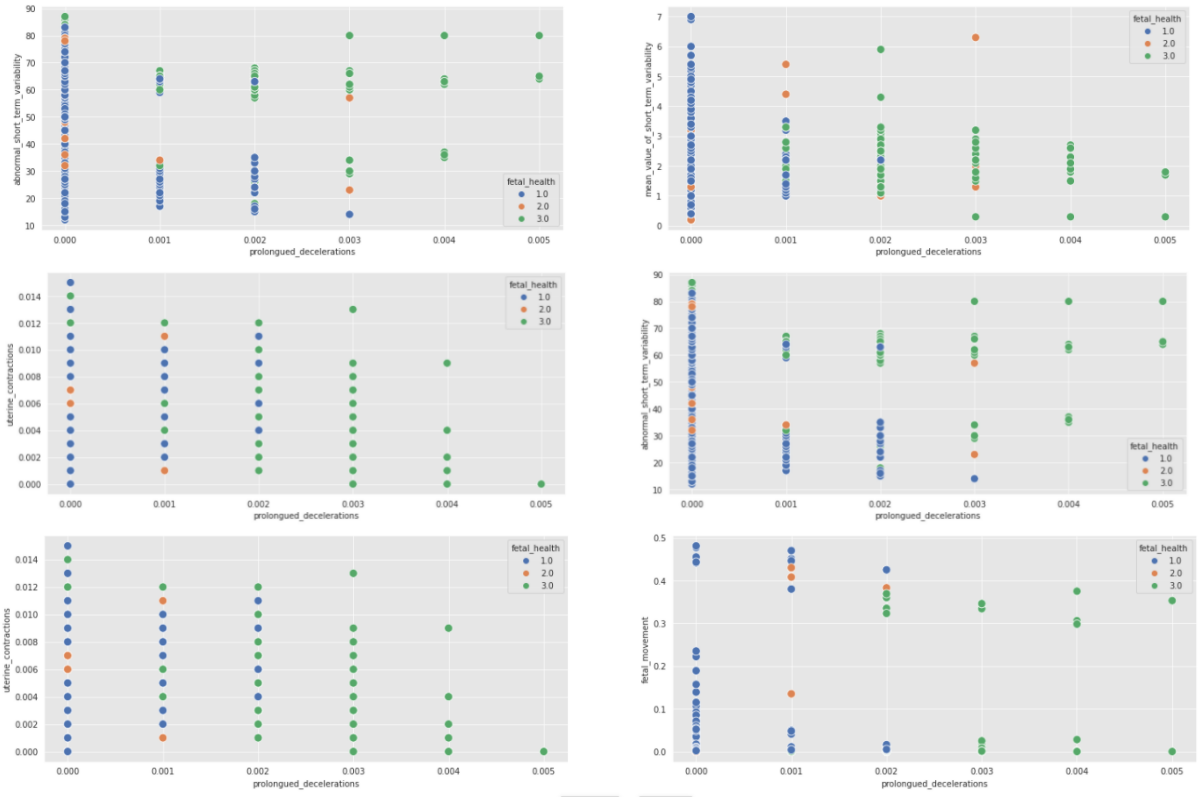
Insights

- No null values, that is great.
- I do not think there are any outliers, in histogram_variance column, there are few values which looks like outliers but not very extreme values. Let us find more about it.



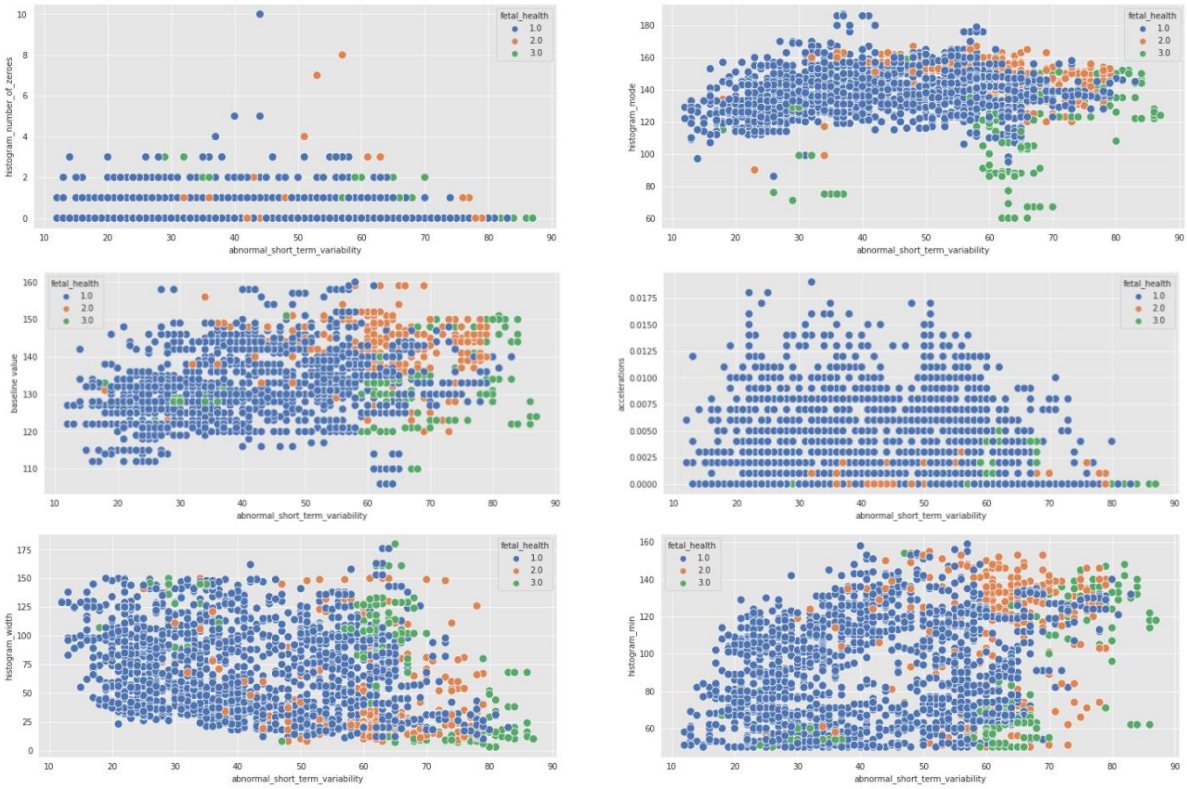
Insights

- Features uterine_contractions, abnormal_short_term_variability and mean_value_of_short_term_variability can be useful classification because I think these can distinguish the class.
- For these features, we see that region of fetal health can be clustered, not completely but to an extent.



Insights

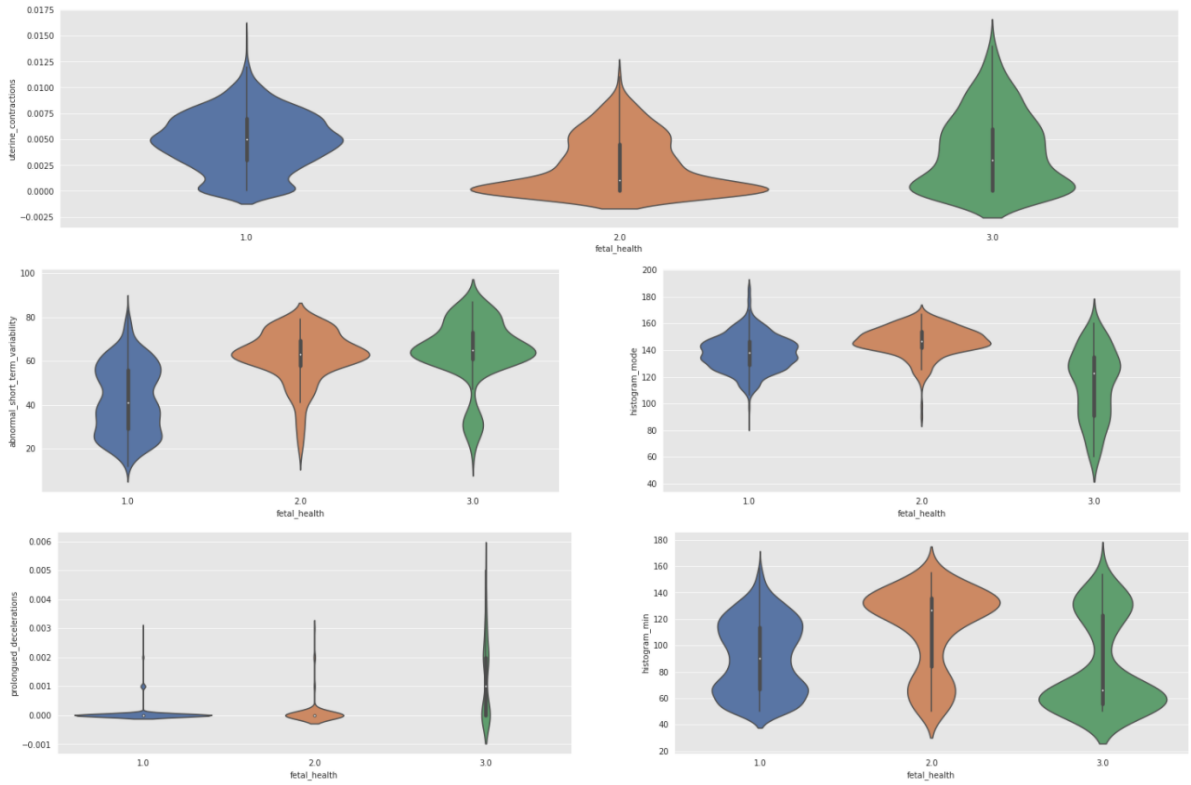
- From the above feature graphs, we see that fetal Pathological can be distinguished, with some error of course, but class Suspect is not easily separated.



Insights

- In the plot, abnormal_short_term_variability vs baseline value there is a small cluster formed after abnormal_short_term_variability value 50 and baseline value greater than 130. Evidently, Suspect cluster is formed.
- Similarly, with the features, abnormal_short_term_variability vs histogram_min, similar cluster is formed.
- Features, abnormal_short_term_variability, histogram_min, histogram_mode, prolongued_decelerations and uterine_contractions can be useful for classification.

Let us investigate distribution of these features:



Insights

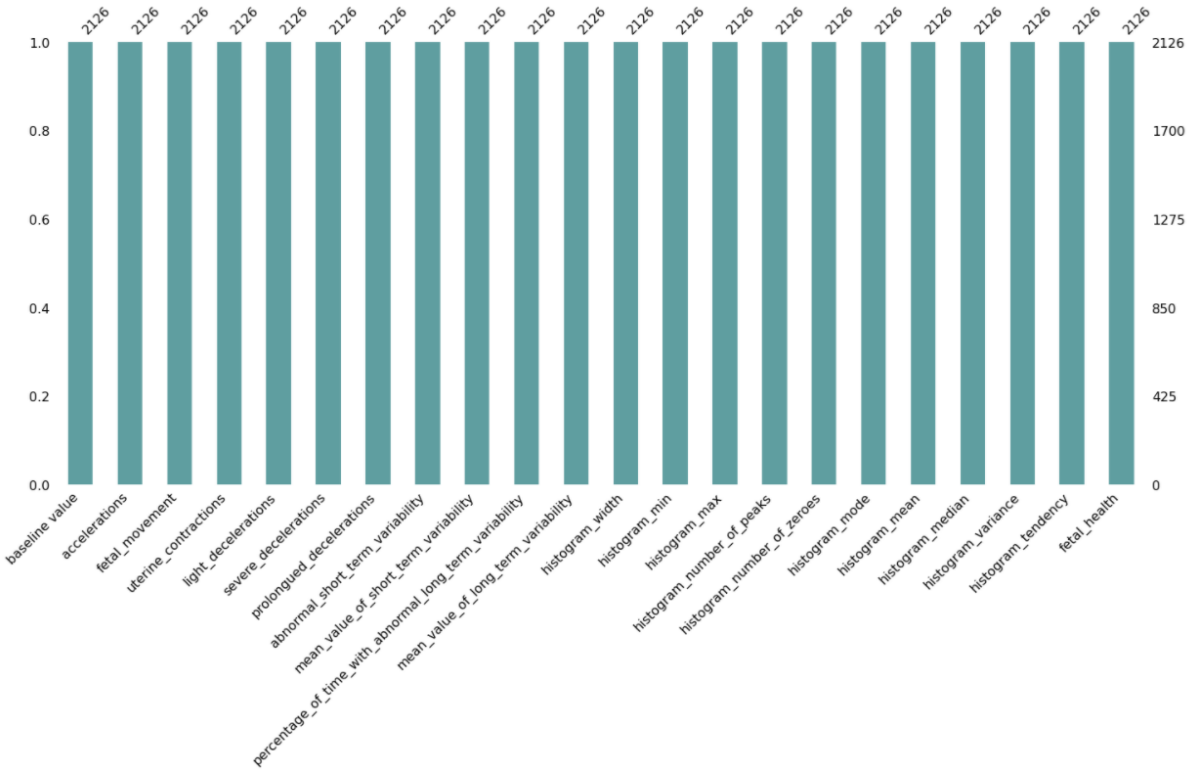
- In feature, abnormal_short_term_variability majority of the Suspected and Pathological points are around 60 and greater than 60.
- In feature, histogram_mode it is little difficult separate Normal and Suspected as they are in almost same distribution.
- In feature, prolongued_decelerations, Normal and Suspected values are around 0, but Pathological are distributed.
- Again, I do not see any specific pattern in histogram_min.
- In feature, uterine_contractions, most of the points in Suspected and Pathological lies around 0, Most important is that there is little proper separation for Suspected class.

2.2 Data Cleaning

Count the missing and null values

Here, it is easy to count the missing and null values. In the case of a real-world dataset, it is very common that some values in the dataset are missing. We represent these missing values as NaN (Not a Number) values. But to build a good machine learning model our dataset should be complete. That is why we use some imputation techniques to replace the NaN values with some probable values.

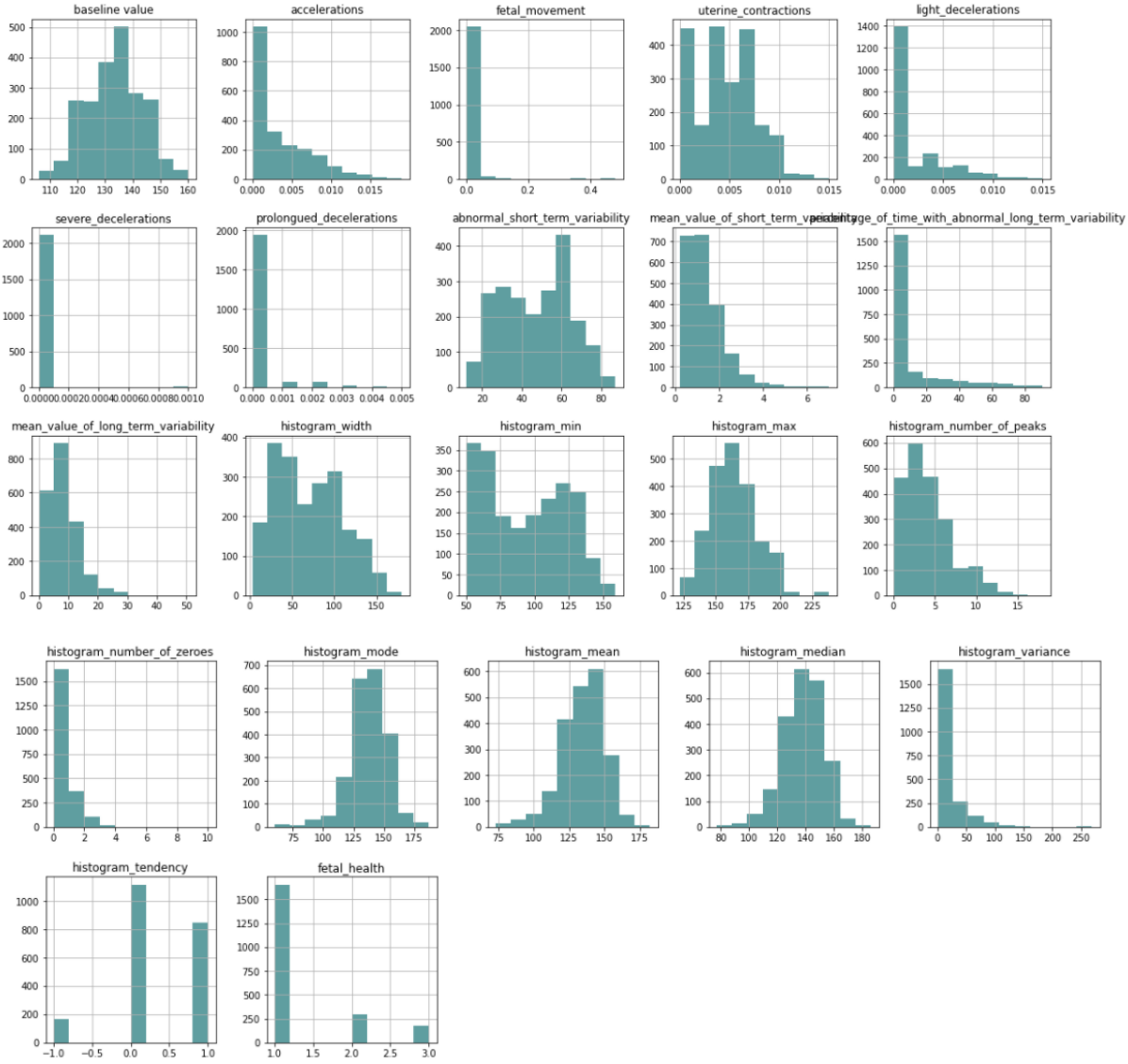
Visualize missing values (NaN) using Missingno Library:



Hence, the dataset is clean, we would not need to clean it further.

2.3 Statistical Analysis to identify key features and correlations

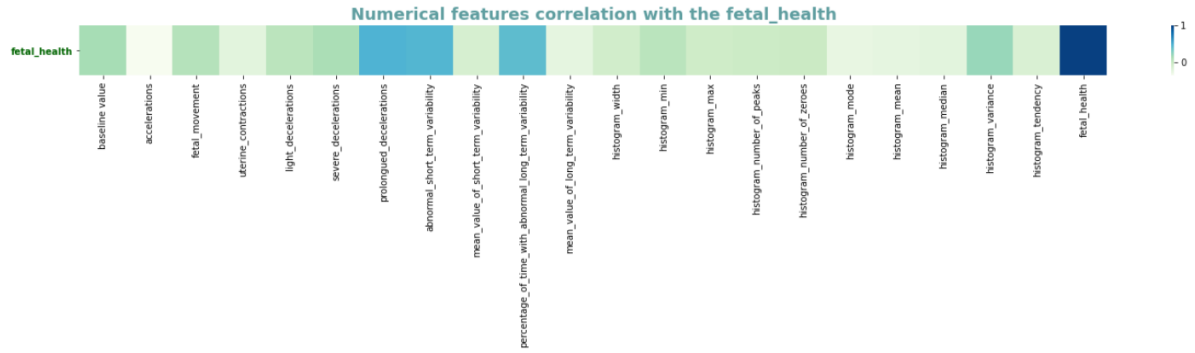
Data visualizations of "fetal_health" column shows us the percentage of fetal health state.



The important things we could learn about the above plot is Skewness. Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. There are three types of skewed distributions. A right (or positive) skewed distribution, left (or negative) skewed distribution, and normal distribution.

- A left-skewed distribution has a long-left tail. Left-skewed distributions are also called negatively skewed distributions. That is because there is a long tail in the negative direction on the number line. The mean is also to the left of the peak.
- A right-skewed distribution has a long right tail. Right-skewed distributions are also called positive-skew distributions. That is because there is a long tail in the positive direction on the number line. The mean is also to the right of the peak.
- The skewness for a normal distribution is zero and looks a bell curve.

Correlation Numeric features with output variable(fetal_health)



	fetal_health
fetal_health	1.000000
prolongued_decelerations	0.484859
abnormal_short_term_variability	0.471191
percentage_of_time_with_abnormal_long_term_variability	0.426146
histogram_variance	0.206630
baseline value	0.148151
severe_decelerations	0.131934
fetal_movement	0.088010
histogram_min	0.063175
light_decelerations	0.058870
histogram_number_of_zeroes	-0.016682
histogram_number_of_peaks	-0.023666
histogram_max	-0.045265
histogram_width	-0.068789
mean_value_of_short_term_variability	-0.103382
histogram_tendency	-0.131976
uterine_contractions	-0.204894
histogram_median	-0.205033
mean_value_of_long_term_variability	-0.226797
histogram_mean	-0.226985

We can see three features: "prolongued_decelerations", "abnormal_short_term_variability", "percentage_of_time_with_abnormal_long_term_variability" have high correlation with the target column (fetal_health).

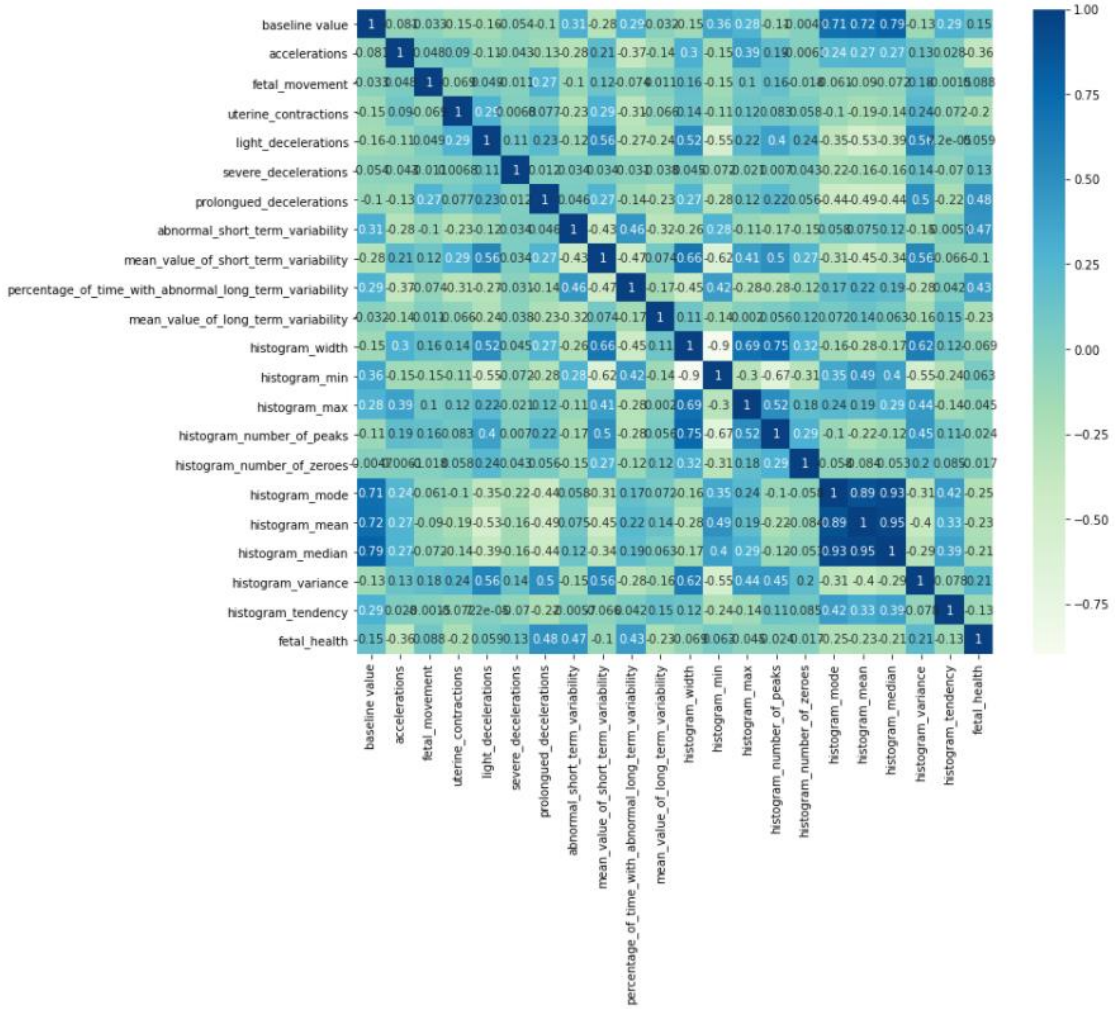
Scatter matrix

A scatter plot matrix is a grid (or matrix) of scatter plots used to visualize bivariate relationships between combinations of variables. Each scatter plot in the matrix visualizes the relationship between a pair of variables, allowing many relationships to be explored in one chart.

It is shown in the jupyter notebook, as here, it was almost impossible to snip and paste the entire image as a whole.

Heatmap

A heat map is a two-dimensional representation of information with the help of colours.



Heat maps can help the user visualize simple or complex information. Correlation heatmaps are ideal for comparing the measurement for each pair of dimension values.

Here, we can clearly see that 'histogram_mode', 'histogram_mean', 'histogram_median' are highly correlated, we might want to exclude them in future.

2.4 Benefits of Feature Engineering

What Is Feature Engineering?

Feature engineering involves leveraging data mining techniques to extract features from raw data along with the use of domain knowledge. Feature engineering is useful to improve the performance of machine learning algorithms and is often considered as applied machine learning.

Features are also referred to as 'variables' or 'attributes' as they affect the output of a process.

Feature engineering involves several processes. Feature selection, construction, transformation, and extraction are some key aspects of feature engineering. Let's understand what each process involves:

- Feature selection involves choosing a set of features from a large collection. Selecting the important features and reducing the size of the feature set makes computation in machine learning and data analytic algorithms more feasible. Feature selection also improves the quality of the output obtained from algorithms.

- Feature transformation involves creating features using existing data by the use of mathematical operations. For example, to ascertain the body type of a person a feature called BMI (Body Mass Index) is needed. If the dataset captures the person's weight and height, BMI can be derived using a mathematical formula.

- Feature construction is the process of developing new features apart from the ones generated in feature transformation, that are appropriate variables of the process under study.

- Feature extraction is a process of reducing the dimensionality of a dataset. Feature extraction involves combining the existing features into new ones thereby reducing the number of features in the dataset. This reduces the amount of data into manageable sizes for algorithms to process, without distorting the original relationships or relevant information.

Why Is Feature Engineering Required?

The intention of feature engineering is to achieve two primary goals:

1. Preparing an input dataset that is compatible with and best fits the machine learning algorithm.
2. Improving the performance of machine learning models.

Common Feature Engineering Techniques Used

- Imputation

One of the most common problems in machine learning is the absence of values in the datasets. The causes of missing values can be due to numerous issues like human error, privacy concern and interruptions in the flow of data among many. Irrespective of the cause, absence of values affects the performance of machine learning algorithms.

Rows with missing values are sometimes dropped by machine learning platforms and some platforms do not accept datasets with missing data. This decreases the performance of the algorithm due to reduced data size. By using the method of Imputation, values are introduced into the dataset that are coherent with the existing values. Although there are many imputation methods, replacing missing values with the median of the column or the maximum value occurred is a common imputation method.

- One-Hot Encoding

This is one of the common encoding methods used in feature engineering. One-hot encoding is a method of assigning binary values (0's and 1's) to values in the columns. In this method, all values above the threshold are converted to 1, while all values equal to or below the threshold are converted as 0. This changes the feature values to a numerical format which is much easier for algorithms to understand without compromising the value of the information and the relationship between the variables and the feature.

- Grouping Operations

In machine learning algorithms, a variable or instance is represented in rows and features are represented in columns. Many datasets rarely fit into the simplistic arrangement of rows and columns as each column has multiple rows of an instance. To handle such cases, data is grouped in such a fashion that every variable is represented by only one row. The intention of grouping operations is to arrive at an aggregation that establishes the most viable relationship with features.

- Log Transformation

A measure of asymmetry in a dataset is known as Skewness, which is defined as the extent to which a given distribution of data varies from a normal distribution. Skewness of data affects the prediction models in ML algorithms. To resolve this, Log Transformations are used to reduce the skewness of data. The less skewed distributions are, the better is the ability of algorithms to interpret patterns.

- Bag of Words

Bag of Words (BoW) is a counting algorithm that evaluates the number of repetitions of a word in a document. This algorithm is useful in identifying similarities and differences in documents for applications like search and document classification.

- Feature Hashing

Feature hashing is an important technique used to scale-up machine learning algorithms by vectorizing features. The technique of feature hashing is commonly used in document classification and sentiment analysis where tokens are converted into integers. Hash values are derived by applying hash functions to features that are used as indices to map data.

Automated Feature Engineering

Automated feature engineering is a new technique that is becoming a standard part of machine learning workflow. The traditional approach is a time consuming, error-prone process and is specific to the problem at hand and must change with every new dataset. Automated feature engineering extracts useful and meaningful features using a framework that can be applied to any problem. This will increase the efficiency of data scientists by helping them spend more time on other elements of machine learning and would enable citizen data scientists to do feature engineering using a framework-based approach.