Fetal Health: Capstone Three

By Morgan Snellgrove

Problem Statement:

One of the goals of the UN and other world health organizations is to reduce under-5 childhood mortality. Observing a reduction in this area is considered a key indicator of human progress. The World Health Organization reported an estimated 5 million deaths in children under the age of 5 in 2020, mostly from preventable and treatable causes. Children in the sub-Saharan region of Africa and the children in Southern Asia have the highest mortality rates in the world. In fact, these regions account for more than 80% of the 5 million deaths in 2020. (Note, only 0.1% of the global death cases from covid were this age group. So, it did not have a significant impact on the under-5 mortality rate.)

Under-5 mortality rates strongly reflect the health and wellbeing of a population's women of reproductive age. It is also a strong indicator of the availability of health care to women and children. Some ways of reducing childhood mortality include prenatal care and screening to prevent and detect hidden conditions. One of the tools that can be extremely useful in detecting hidden conditions is a CTG, cardiotocography, exam. However, in areas where health care professionals are few, there aren't enough doctors to perform and read the results of such an exam. A model that could help identify the suspect and pathological fetal health cases would be a great asset to medical professionals in these areas.

Goals:

Build a model that could identify the category of fetal health with greater than 80% accuracy.

The main category of interest is class 3: pathological. We want a greater than 90% recall on this category.

Identify which CTG features were most important in making these classifications.

The Data:

This model was built using data from:

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

We have 2126 observations in our dataset. Of those:

78% are class 1: Normal,

14% are class 2: Suspect,

and 8% are class 3: Pathological.

Terminology:

To begin, I decided to learn a little more about the terminology used in the column names of the dataset. This would help me assess if my findings made sense or not. The following terminology is useful when discussing CTG results:

1) baseline value = the average heart rate of the fetus within a 10-minute window. Normal fetal heart rate is between 110 - 160 bpm. The fetal heart rate is shown in the top line on the CTG output.

2) fetal tachycardia = a baseline fetal heart rate over 160.

3) fetal bradycardia = a baseline fetal heart rate under 110. Sever prolonged bradycardia is fetal heart rate less than 80bpm for over 3 minutes.

4) uterine contractions = Shown on the bottom line in the CTG output. A contraction appears as a peak on this line. The number of contractions present per 10-minute window is measured.

5) variability = the variation of fetal heart rate from one beat to the next. Normal variability is 5 - 25 bpm. Variability is calculated by assessing how much the peaks and troughs deviate from the baseline rate.

6) abnormal variability = variability of less than 5bpm for 50 minutes, more than 25bpm for 25 minutes, or sinusoidal (0 variability, sine shape)

7) accelerations = abrupt increase in baseline rate of more than 15bpm for longer than 15 seconds. The presence of accelerations is reassuring. Accelerations occurring alongside uterine contractions is an indicator of a healthy fetus.

8) decelerations = abrupt decrease in baseline rate of more than 15bpm for longer than 15 seconds. When a fetus experiences hypoxic stress (low oxygen) it reduces it's heart rate to maintain myocardial oxygenation (oxygen needed by the heart) and perfusion (sending blood/fluid to the rest of the body). When adults experience hypoxic stress, they will breathe deeper and faster to bring in more oxygen. A fetus cannot do this so they adjust their heart rate instead.

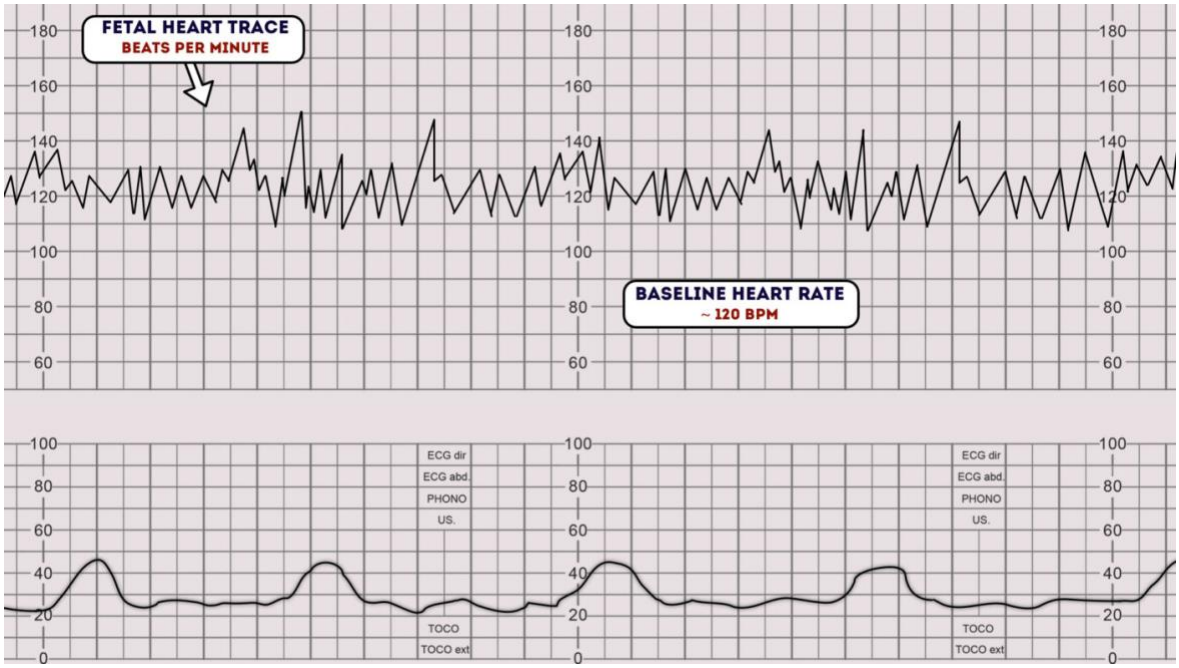
9) early decelerations = a deceleration that starts when a uterine contraction begins and recovers when the contraction stops. This is due to increased pressure from the contraction. This type is considered physiological and not pathological.

10) variable decelerations = a rapid fall in baseline rate with a variable recovery. Duration varies and they may not be associated with contractions. Common during labor. The presence of persistent variable decelerations indicates the need for close monitoring.

11) late decelerations = begin at the peak of the uterine contraction and recover after the contraction ends. This type of deceleration indicates there is insufficient blood flow to the uterus and placenta. As a result, blood flow to the fetus is significantly reduced causing fetal hypoxia and acidosis.

12) prolonged deceleration = a deceleration that lasts more than 2 minutes. If it lasts longer than 3 minutes it is immediately classed as abnormal.

Example CTG Output:



EDA:

After importing the data and the necessary libraries, I called .info to see my datatypes and if we had nulls in any of the columns. All 22 columns were float64 and showed 2126 non-null entries.

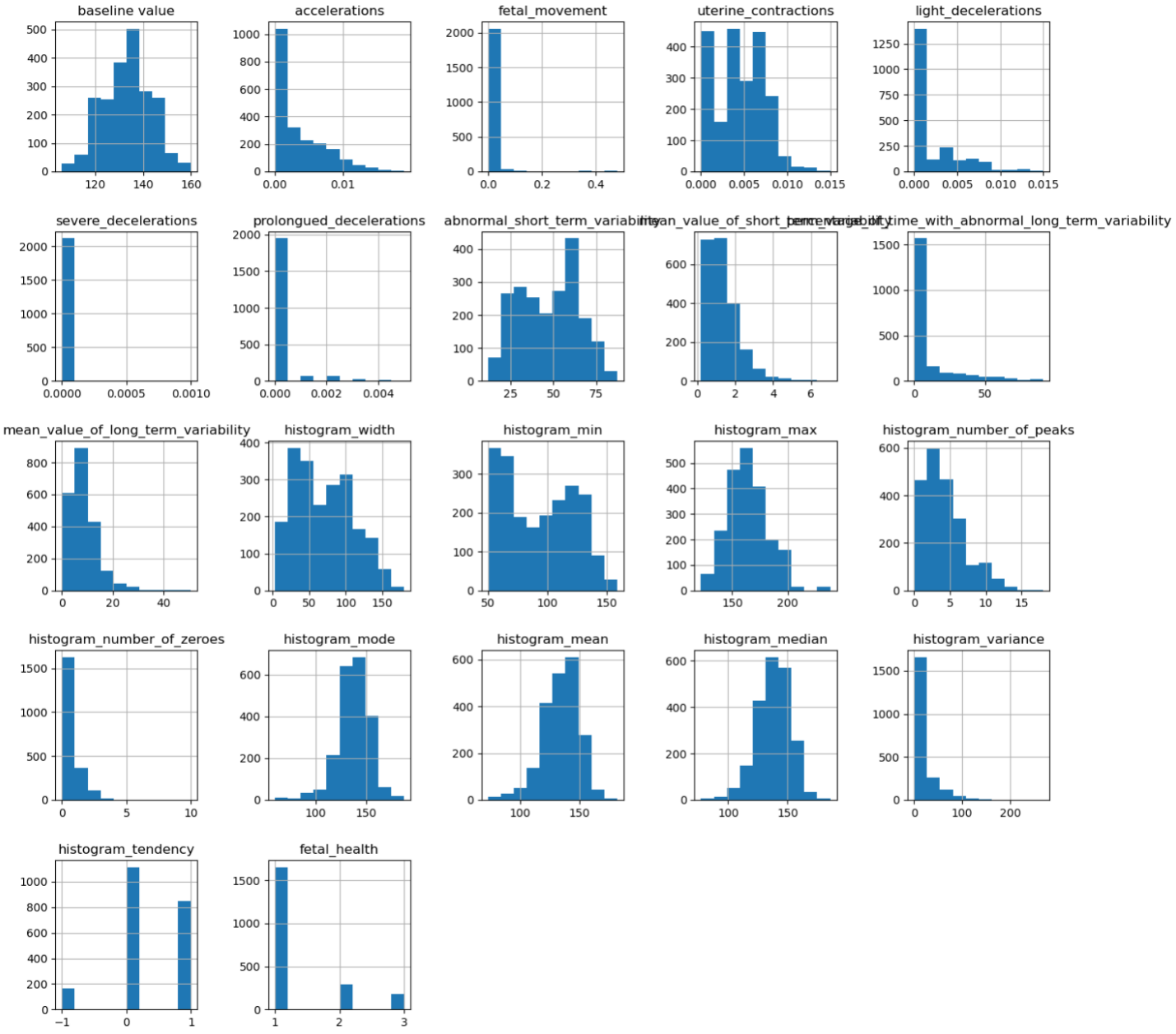
Next, I used .describe to see the summary statistics for each column. I noted the minimum and maximum of the column baseline value was 106-160. From the terminology, we learned that the healthy range of fetal heart rate, FHR, is 110-160. So, we have some low values in this column.

I could also see that the columns had very different scale from each other. Many had maximums that were less than 1 while others had maximums near 250.

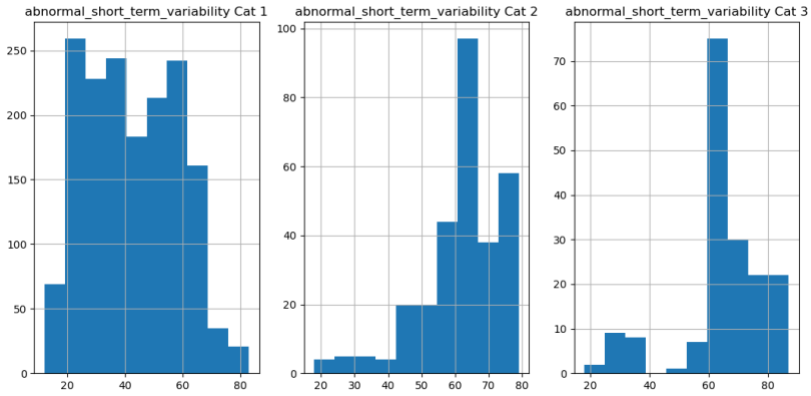
I also saw that we had several columns that were measurements taken from a histogram. After some research, I found that these histograms showed the distributions of the individual FHR readings. (A fetus' heart rate has a lot of variability.)

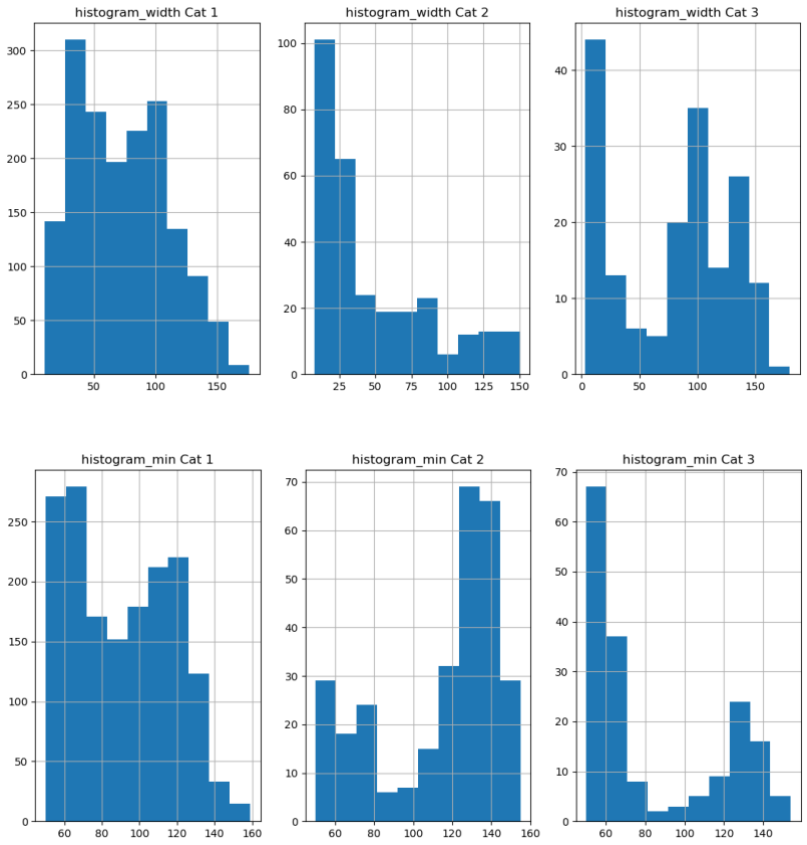
Next, I made a count plot of the target, fetal_health. This showed we had 1655 examples of class 1, 295 of class 2, and 176 of class 3. So, I knew we were working with unbalanced data.

Then I looked at the distributions for each column. Baseline value of FHR was very close to normally distributed, as were histogram_max, histogram_mean, histogram_median, and histogram_mode. I also noticed that three distributions had more than one peak: abnormal_short_term_variability, histogram_width, and histogram_min.



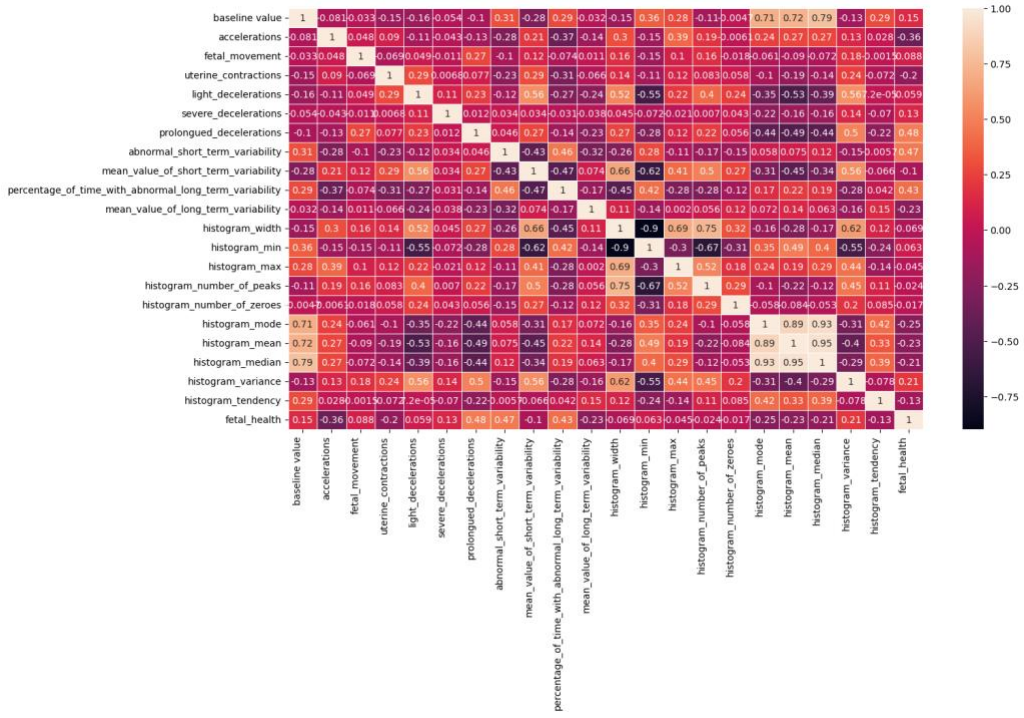
I decided to subset the data by the target, and then look at each column's distribution. Here is what I found for the three bimodal columns:





We continue to see the bimodal distribution in histogram_min for all three categories. By looking at the description of the full dataset, we saw that histogram_width had a minimum of 3. Calling .describe on each target subset revealed that all of the histogram_width = 3 examples were in category 3. This is why that distribution is still bimodal.

Next, I looked at the correlation heat map for the full dataset.



We can see strong negative correlation between the features histogram_width and histogram_min. (These were two of our bimodal distributions.) We also see really strong positive correlation between histogram_mean, histogram_median, and histogram_mode.

As far as correlation with the target, we don't see anything very strong. All correlation with the target was weak to moderate, with some columns showing close to zero correlation. The strongest positive correlation with the target is with prolonged_decelerations and abnormal_short_term_variability. This makes sense because both were mentioned as negative signs of fetal health in the terminology section. The strongest negative correlation with the target is with accelerations, uterine_contractions, mean_value_of_long_term_variability, histogram_mean, histogram_median, and histogram_mode. Once again, these make sense from the terminology section. The presence of accelerations is reassuring and looking at the distributions for mean_value_of_long_term_variability we do see more lower values for classes 2 and 3.

Preprocessing:

To prepare the data for modeling, I split it into a feature dataframe and a target series. The feature dataframe, X, contained all columns except for 'fetal_health'. The target series, y, is just the column 'fetal_health'.

I applied the train test split function to X and y with the stratify argument set to y and the random state set to 42.

Then I fit the Standard Scaler to X train and transformed both X train and X test. I did not apply it to y test or train.

Lastly, I went back and included the column names to help interpret the results.

After these steps the data was ready for modeling.

Modeling:

To have something with which to compare our models, I built a dummy classifier. I used the "most frequent" strategy so that it always predicted class 1. The overall accuracy of this model was 0.78. This is because 78% of our dataset is class 1.

The first real model I tried was a Decision Tree. I built a few versions: Entropy with no max depth, Gini with no max depth, Entropy with max depth of 3, Gini with max depth of 3. Of the four models, the highest accuracy score of 0.91 came from Entropy with no max depth. The model with the highest class 3 recall, 0.86, was Gini no max depth.

Next, I tried a Random Forest model. Out of the box, this model beat the Decision Trees. Its accuracy was 0.93 and its class 3 recall was 0.89.

The last model I tried was a Gradient Boosting Classifier. This model, without any tuning, gave an accuracy of 0.93, and a class 3 recall of 0.93. This was looking like the most promising model.

Now that I had tried some out-of-the-box models, I wanted to see how they improved with some hyperparameter tuning. Since the dataset wasn't that big, I chose GridSearchCV to test out some values. I did GridSearchCV on all three of the model types. The best parameters for each type of model were:

Decision Tree: Entropy with Max Depth = 7

Random Forest: Entropy with Max Depth = 13

Gradient Boosting: Squared Error with Learning Rate = 0.2

Then I built a new model for each type with these parameters. The best model of this group was the Gradient Boosting Classifier. Its overall accuracy was 0.94 and its class 3 recall was 0.93.

Creating a Binary Target:

Looking back over all the models, each one seemed to struggle with identifying class 2. Every model had its lowest recall on class 2. Since this category is also important for catching medical issues, I thought it might be helpful to re-bin the target values.

To do this, I copied the original full dataset and created a new column, binary_target, that was 1 if fetal health equaled 2 or 3 and was 0 if fetal health equaled 1. I followed the same preprocessing steps with this new dataset, letting X be all columns except fetal health and binary target and setting y as the column binary target. I applied train test split (with the same arguments) and scaled X train and X test.

I went straight into GridSearchCV on the same model types as before. Once again, the best metrics came from Gradient Boosting Classifier. This model had an accuracy of 0.96 and a class 1 (combined classes 2 and 3) recall of 0.88. The recall fell somewhere between the recall values for classes 2 and 3 from the multiclass target model, but the overall accuracy had increased slightly.

Final Model Selected:

In the end, the Multiclass Gradient Boosting Classifier, met all our goals, while the Binary Gradient Boosting Model fell short in class 3 recall. I would recommend the multiclass approach with the final model from that section.

Feature Importances:

The top three features for the multiclass GB model were: abnormal_short_term_variability, mean_value_of_short_term_variability, and percentage_of_time_with_abnormal_long_term_variability.

The top three features for the binary GB model were: mean_value_of_short_term_variability, histogram_mean, and abnormal_short_term_variability.