DATA699: Capstone Research Project

Fetal Health Classification

Aina Ahanmisi

Table of Contents

Introduction

Every year, according to the Center for Disease Control and Prevention, about one million fetal deaths occur in the United States and almost three million in the world (DeSisto et al. 2021). The fetal mortality rate has been declining slowly since 1990 but rates have started to plateau starting 2003 and have not moved since. The number of fetal deaths in the United States have been at a steady rate for years now and has begun increasing in some states. (Centers for Disease Control and Prevention 2015). The reduction of fetal death is an important goal because it has many beneficial effects for all. Preventing fetal deaths can also assist with preventing maternal mortality and can also help with new discoveries on a variety of different things like on what new health conditions and risk factors that may be disrupting the pregnancy term. Reducing the number of fetal deaths will in fact be a great help in human progression by advancing scientifically and offering proper world evolution (The Global Goals For Sustainable Development 2021).

1.1 Problem Identification

The objective of this project is to assist in the goal by properly identifying the status of a fetus early enough to reverse the causes leading to death through data exploration and classification. The target audiences of this project are professionals in the Center for Disease Control and Prevention, World Health Organization, medical practitioners and scientists. These target audiences are groups of people who are able to make decisions and budget for what studies deserve funds to continue with further studies. This audience would care for fetal health because the research aids the global goal to help with human progression. This project can also target regular civilians who are mothers and fathers, more specifically, people with children. I believe mothers who have experienced fetal deaths would like to know this information or find comfort in research being done to help reduce this issue.

Data Description

The data set that will be used for this project is from Kaggle. There are a total of 2126 women in this Fetal Health Classification data set. The women in this data set have already been classified into a specified category of either normal, suspect or pathological, based on the status of their child. Normal (1) is considered to be a healthy fetus because it aligns with a normal status of a healthy fetus. The suspect (2) category is considered to be on the borderline of an unhealthy and healthy fetus and has a possible risk of death. The pathological (3) category is considered to be an extreme case, with the fetus being severely unhealthy and at high risk of death.

The features of this data set are medical measurements extracted from cardiotocograms that the women have taken and have been classified by expert obstetricians. There are a total of 22 attributes in the Fetal Health Classification data set. The table in Table 1 condenses the 22 attributes into categories below to better briefly describe what each represents from the data set. All attributes are measurements of the fetus and not the mother.

Table 1: Description of Data

Features	Description
Heart Rate (baseline value)	Measurement of the fetal heart rate where 135 beats per minute is considered to be the normal baseline fetal health rate
Heart Acceleration (accelerations)	Measurement of sudden acceleration of the fetal heart rate. Abrupt acceleration is a good sign of a normal fetus, because this is heavily associated with healthy fetal movement.
Movement (fetal_movement)	Measure of fetal movements per second, all fetal movement is good. There is nothing bad with an active baby because such movement promotes healthy bone and joint development (Bogle, 2020).
Uterine Contractions (uterine_contractions)	Measure of uterine contractions per second. This measurement can help with infection possibilities or overall well-being of the fetus.
Decelerations (light_decelerations, severe_decelerations, prolongued_decelerations)	Broken into three attributes which are light, severe and prolonged. Decelerations are sudden but very temporary drops in the fetal heart rate. Light decelerations can be normal but severe or prolonged decelerations are not a good sign. These attributes are also correlated to uterine contractions.
Variability Calculations (abnormal_short_term_variability, mean_value_of_short_term_variability, percentage_of_time_with_abnormal_long_term_variability, mean_value_of_long_term_variability)	Consist of four attributes that are all either measurements or calculations of the variations of the fetal heart rate from beat to beat (Bailey 2009). Calculations include abnormality and mean value of the short term. And percentage of abnormality and mean

	value of long term.
Histogram (histogram_width, histogram_min, histogram_max, histogram_number_of_peaks, histogram_number_of_zeroes, histogram_mode, histogram_mean, histogram_median, histogram_variance, histogram_tendency)	Consist of ten attributes that all are used to form a histogram. These attributes are calculated using all values from the cardiotocography report. The calculations include the width, minimum, maximum, number of peaks and zeroes, mode, mean, median, variance and tendency.
Fetal Health (fetal_health)	Each woman out of the 2126 are classified into the fetal health categories based on their report (Campos et al. 2000)

Methodology

In this section, the exploration analysis is reported and the classification analysis is addressed using different classifiers that include linear regression, random forest and decision tree models.

3.1 Data Preparation and Exploration

Before proceeding with exploring the data, I attempted to clean it first by checking for missing values in the data set by searching for anything with NULL or NA values. "Missing data are values that are not recorded in a dat set" (Alam 2020). We check for missing data because it can potentially cause skewness in the data and provide results that could be inaccurate. So it is always best to work on data with no missing values. This data set, thankfully does not have any missing values, so we did not need to remove anything from the data.

While exploring the data and looking at the spread distribution of values across the categories of the response variable which is represented in Figure 1 below, I realized that there is a significant imbalance between each of the categories. The normal category contains the majority of the data at about 78% while suspect and pathological are at 14% and 8%. The imbalance of the response variable means that the distribution across the classes are biased or skewed. To fix this imbalance, I decided to upsample the data. "Upsampling is the process of randomly duplicating observations from the minority class in order to reinforce its signal" (Elite Data Science 2019). Figure 1 and Table 3 show the distribution and number of classes before being upsampled. The suspect and pathological classes of the response variable were upsampled to equal the number of normal categories, making all categories 1655 samples each which brings the new total number of samples to be 4965. A brief preview and summary of the new data is reported in Table 2.
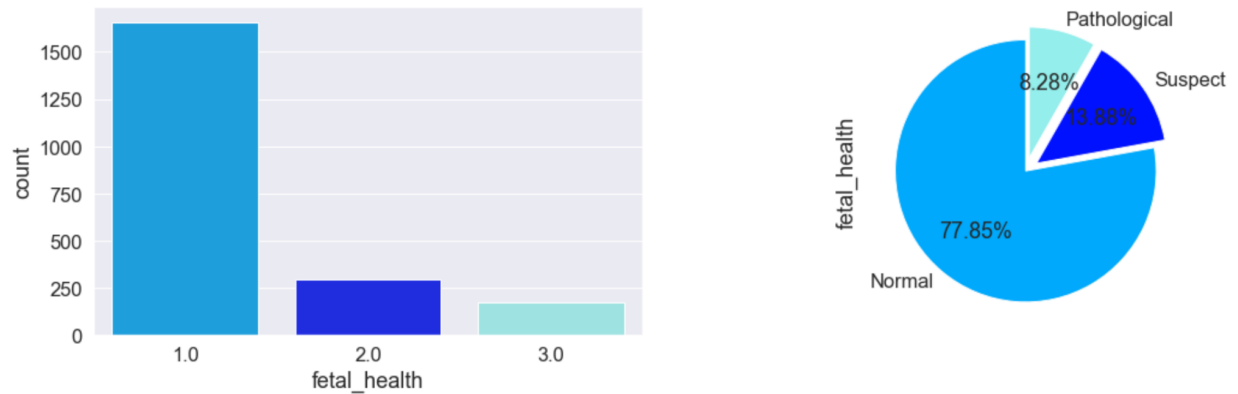
Figure 1: Distribution of Response Variable

Table 2: Summary of Data Set

	count	mean	std	min	25%	50%	75%	max
baseline value	4965.0	135.084391	10.075937	106.0	128.0	134.000	143.000	160.000
accelerations	4965.0	0.001547	0.002972	0.0	0.0	0.000	0.002	0.019
fetal_movement	4965.0	0.013575	0.059262	0.0	0.0	0.000	0.003	0.481
uterine_contractions	4965.0	0.003666	0.003185	0.0	0.0	0.003	0.006	0.015
light_decelerations	4965.0	0.002091	0.003326	0.0	0.0	0.000	0.003	0.015
severe_decelerations	4965.0	0.000012	0.000109	0.0	0.0	0.000	0.000	0.001
prolongued_decelerations	4965.0	0.000481	0.001025	0.0	0.0	0.000	0.000	0.005
abnormal_short_term_variability	4965.0	56.444713	17.011254	12.0	45.0	61.000	67.000	87.000
mean_value_of_short_term_variability	4965.0	1.218348	1.004065	0.2	0.4	0.900	1.800	7.000
percentage_of_time_with_abnormal_long_term_variability	4965.0	18.849748	25.716783	0.0	0.0	4.000	33.000	91.000
mean_value_of_long_term_variability	4965.0	6.762558	5.189628	0.0	3.6	6.500	9.300	50.700
histogram_width	4965.0	67.379456	43.703406	3.0	26.0	64.000	105.000	180.000
histogram_min	4965.0	95.710775	33.144352	50.0	63.0	96.000	128.000	159.000
histogram_max	4965.0	163.090232	18.710513	122.0	151.0	160.000	174.000	238.000
histogram_number_of_peaks	4965.0	3.942397	3.114764	0.0	1.0	3.000	6.000	18.000
histogram_number_of_zeroes	4965.0	0.309970	0.729260	0.0	0.0	0.000	0.000	10.000
histogram_mode	4965.0	132.861229	22.762797	60.0	125.0	139.000	148.000	187.000
histogram_mean	4965.0	130.792346	20.937750	73.0	120.0	136.000	146.000	182.000
histogram_median	4965.0	134.754683	18.741784	77.0	123.0	139.000	149.000	186.000
histogram_variance	4965.0	25.750655	43.507799	0.0	1.0	4.000	35.000	269.000
histogram_tendency	4965.0	0.232226	0.666217	-1.0	0.0	0.000	1.000	1.000
fetal_health	4965.0	2.000000	0.816579	1.0	1.0	2.000	3.000	3.000

Table 3: Count of Response Variable before Upsampling

Distribution of Classifier	
1 (Normal)	1655
2 (Suspect)	295
3 (Pathological)	176

To further explore the data, I looked at the frequency distributions of values for each feature (Figure 2). Looking at the data let me know whether the data needs to be scaled or not so more dominant features will not overpower the others for the analysis. After looking at each distribution, I found that there is indeed a lot of skewness and outliers.
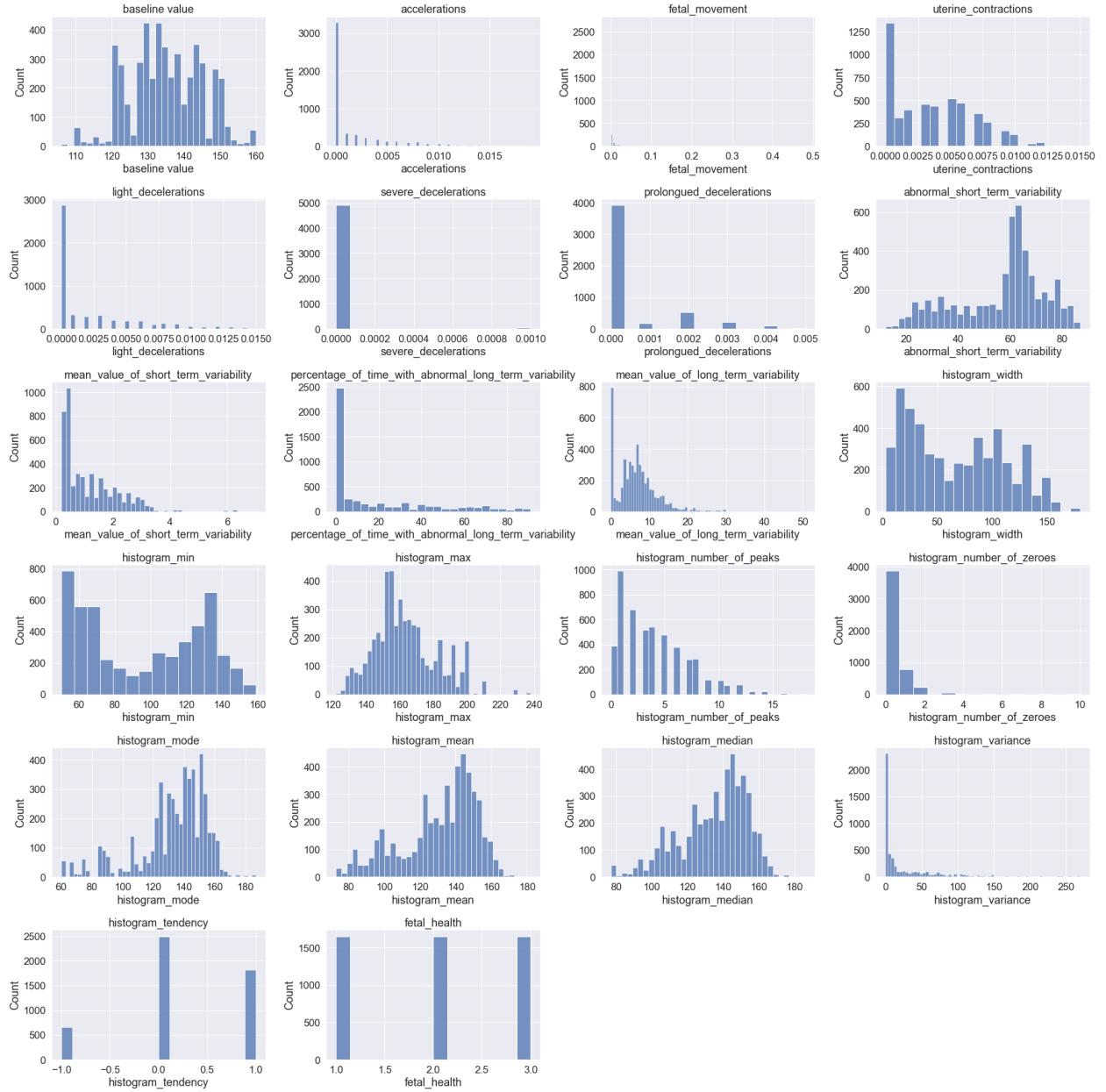
Figure 2: Frequency Distributions of Features

Next, I explored the correlations of the features with the response variable within the data set. I analyzed the correlations by looking at a heatmap of the data and the relationship between the predictor versus the response variable which is represented in Figure 3 and 4 below. Based on the heatmap and relationship between the variables, the highest correlation to fetal health is abnormal short term variability at 53%. For further analysis, I used all features at 40% or higher. In Table 4, I have parsed out the variables that I chose to use for further analysis. The features that were used are accelerations, mean value of long term variability, histogram mode, histogram mean and histogram median which are negatively correlated to fetal health and prolonged

decelerations and abnormal short term variability which are positively correlated to fetal health. These features showed to have higher counts when looking at them versus the response variable.
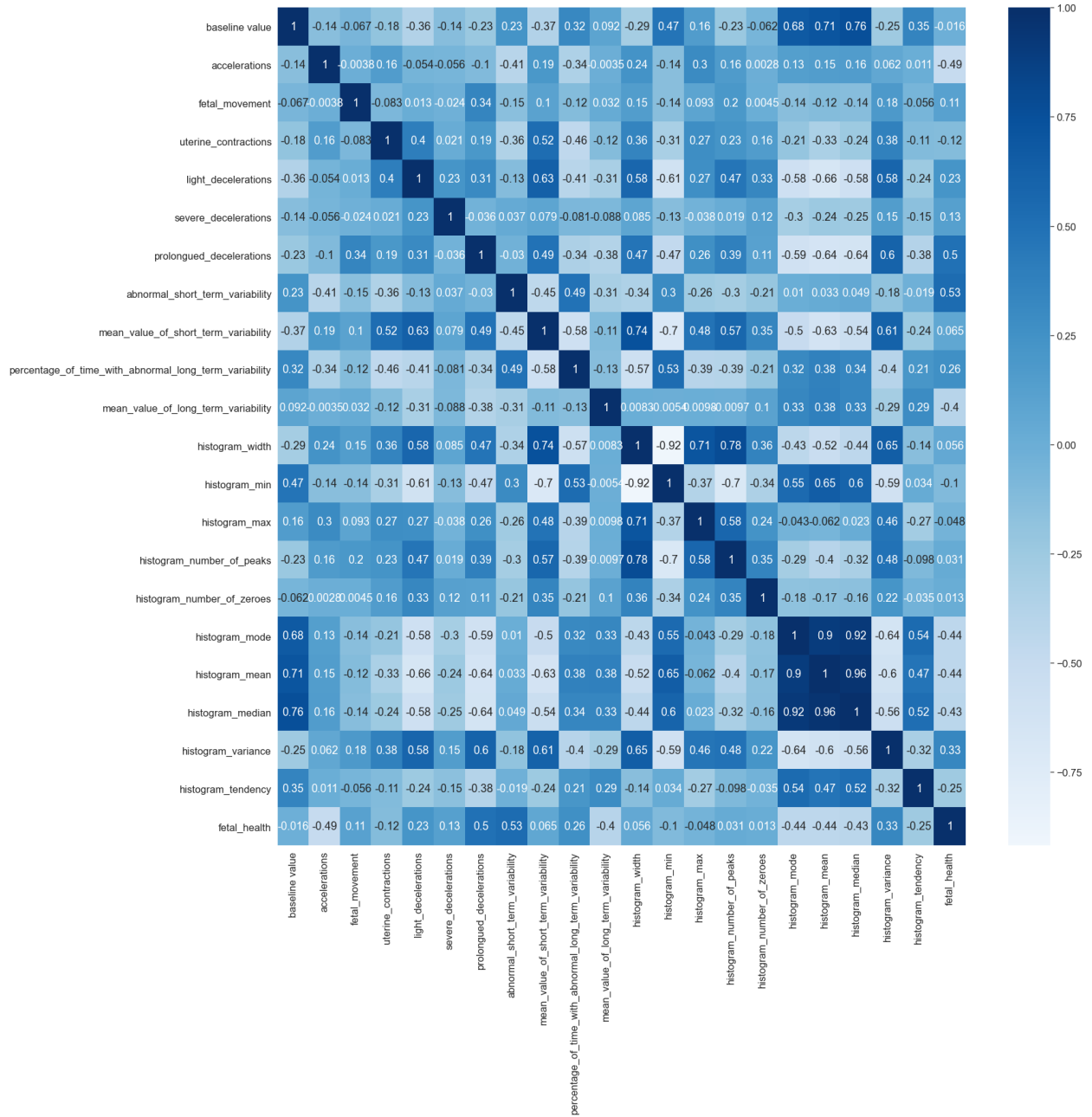

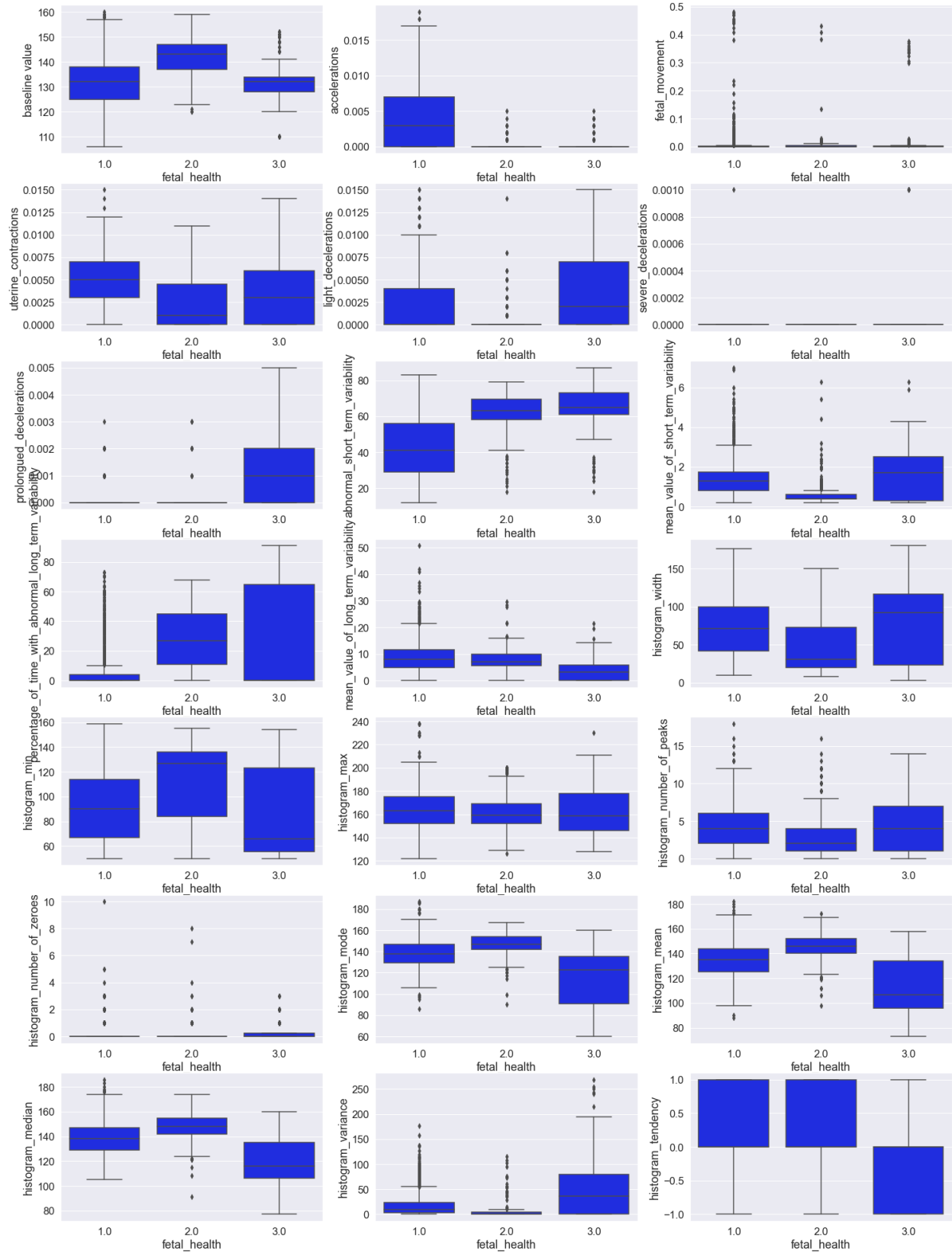**Figure 3: Correlation of Features with Fetal Health**

Figure 4: Predictor versus Response Relationship

Table 4: Correlations Higher than 40%

Feature	Correlation Percent
accelerations	-0.49
prolongued_decelerations	0.5
abnormal_short_term_variability	0.53
mean_value_of_long_term_variability	-0.4
histogram_mode	-0.44
histogram_mean	-0.44
histogram_median	-0.43

Before proceeding with modeling the data, I wanted to closely explore the spread distributions of classes with each of the features that I chose to use for my analysis and discuss some of them (Figure 5).

Accelerations

This distribution mostly contains the suspect and pathological classes. "Accelerations are short-term rises in the heart rate of at least 15 beats per minute" (Wilson 2018) which on average last for about 15 seconds. Accelerations are actually very common, healthy and represent a good oxygen supply. Although accelerations are good, the majority of the instances of each class had measurements of 0.

Prolonged Deceleration

This distribution mostly contains normal and suspect classes. Prolonged Decelerations is when the fetal health rate decreases below the baseline of 15 beats per minute or more, lasting at least two minutes but less than ten minutes from onset to return to baseline (UpToDate 2021). Prolonged decelerations can mean many things but one common indication is cord compression from oligohydramnios (Dildy 2005). The majority of the instances of each class in this data set also had measurements of 0.

Abnormal Short Term Variability

This distribution has a good amount of all three classes. Short term variability is the oscillation of the fetal heart rate around the baseline in amplitude of five to ten beats per minute (Sweha et al. 1999). The classes within this distribution look to be distributed evenly.

Mean Value of Long Term Variability

This distribution is negatively skewed and contains mostly suspect and pathological classes. Long Term Variability is a "slower oscillation in the heart rate and has a frequency of three to ten cycles per minute and an amplitude of 10 to 25 beats per minute (Sweha et al. 1999).

Histogram Mode, Mean and Median

These distributions are slightly negatively skewed and the instances for the suspect class have the highest measurements followed by normal and suspect.
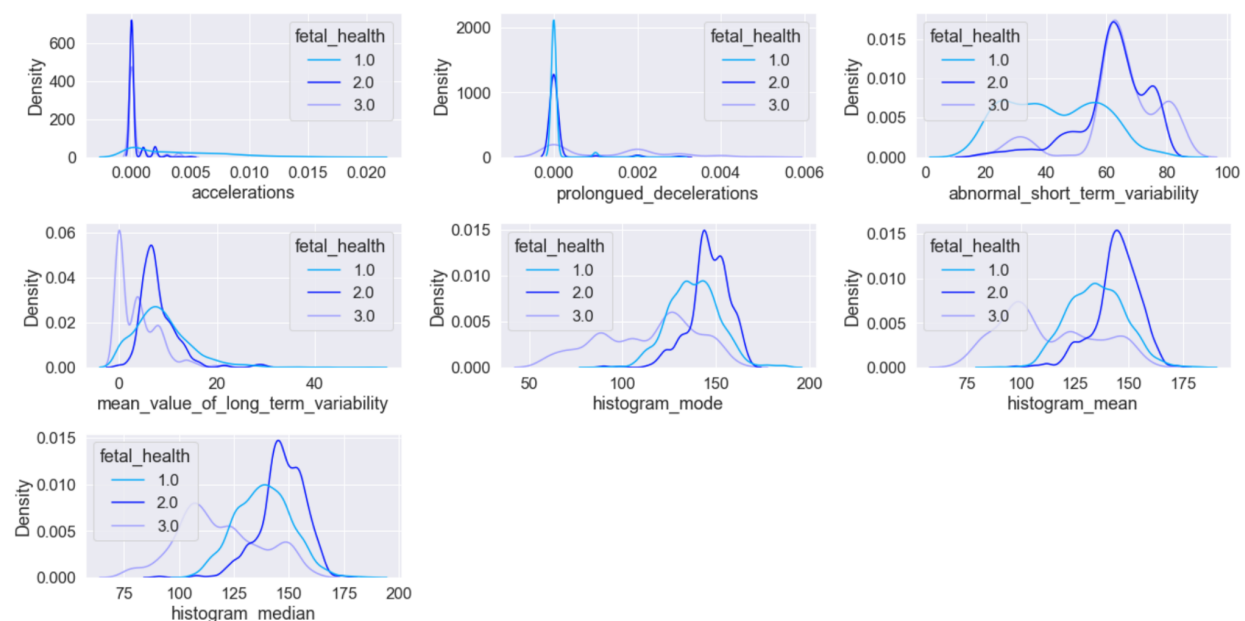

Figure 5: Distributions of Correlated Features versus Fetal Health

3.2 Classifier Models

To get started on modeling, I first scaled the data by using the StandardScaler function (Figure 6) and then split the data to train and test using train_test_split function with a split of 70% training and 30% testing.
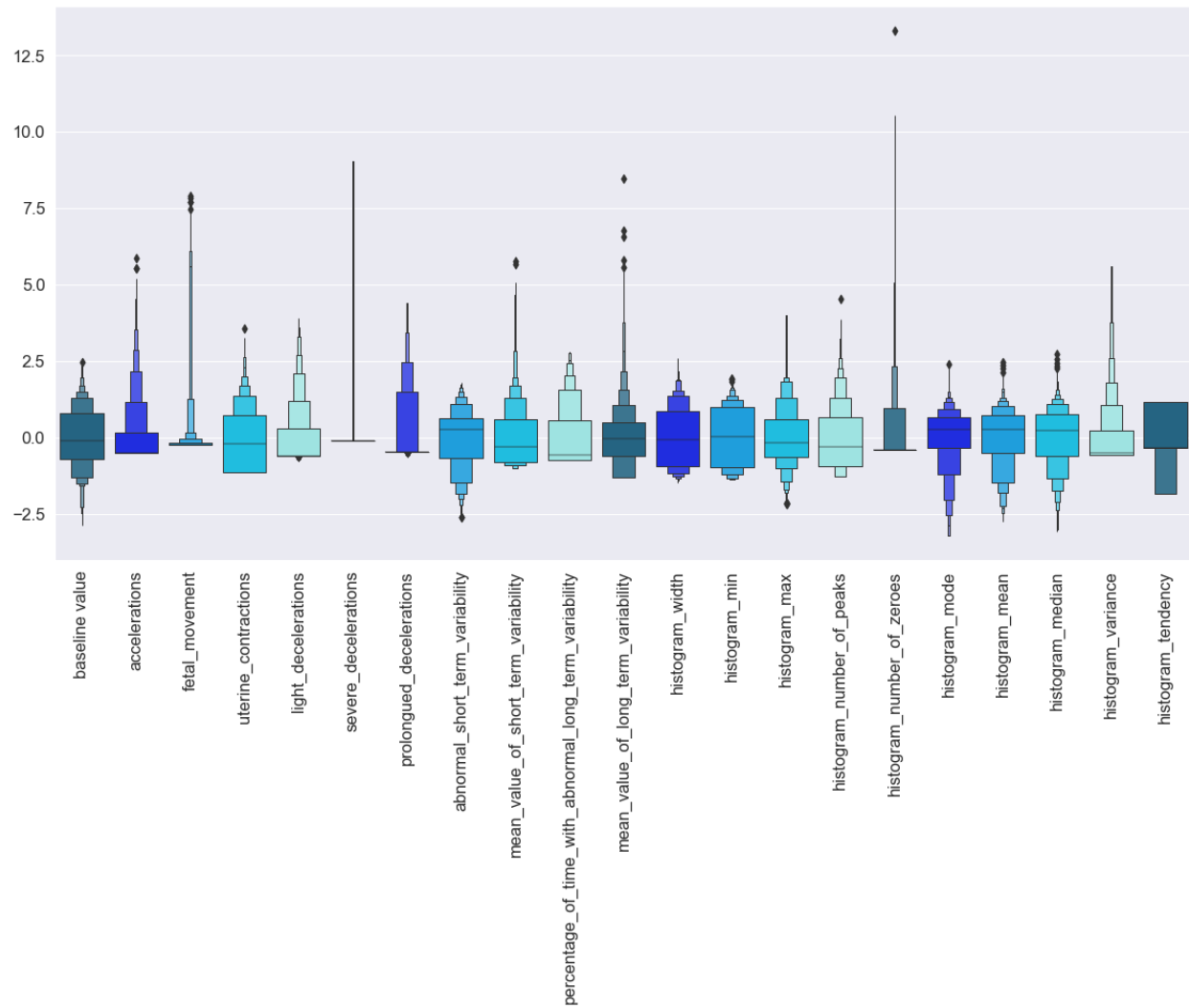
Figure 6: Box Plot of Scaled Data

3.2.1 Model 1: Linear Regression

The classifier used to build model 1 is the linear regression classifier instantiated with a 'liblinear' solver and the most correlated features chosen were passed into the classifier (Gross et al. 2003). Using these features, the model computed an accuracy of 81% and an error rate of 19%. Although this is a good baseline score, I believe the score can be better.

Table 5: Linear Regression Scores

	precision	recall	f1-score	support
1.0	0.91	0.80	0.85	531
2.0	0.71	0.83	0.77	474
3.0	0.81	0.79	0.80	485
accuracy			0.81	1490
macro avg	0.81	0.81	0.81	1490
weighted avg	0.82	0.81	0.81	1490
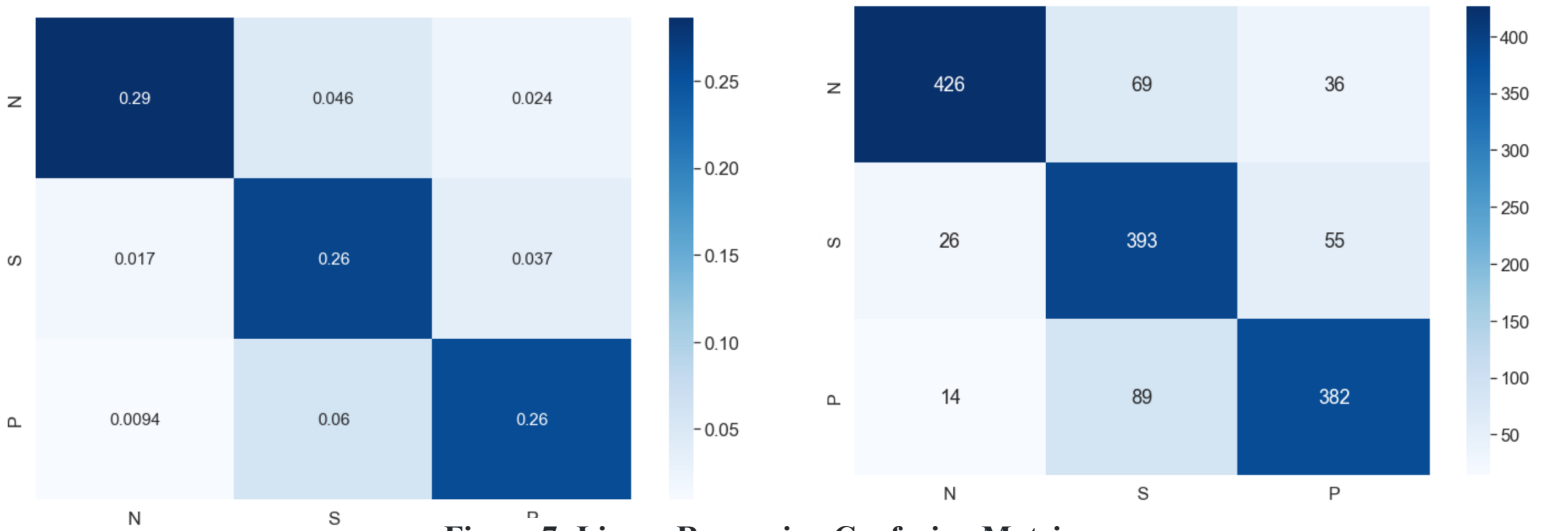
Figure 7: Linear Regression Confusion Matrix

3.2.2 Model 2: Random Forest

The classifier used to build model 2 is the Random Forest classifier which is an ensemble estimator based on multiple random decision trees. The model works by random selection of features and running them through several iterations of smaller decision trees and aggregating the results (Oshiro et al. 2012). Using the sklearn library, the following parameters were passed into the classifier, "criterion = 'gini',n_estimators=100,max_depth=4,random_state=33". With the parsed features with the highest correlations, the model computed an accuracy of 85% and an error rate of 15%.

Table 6: Random Forest Scores

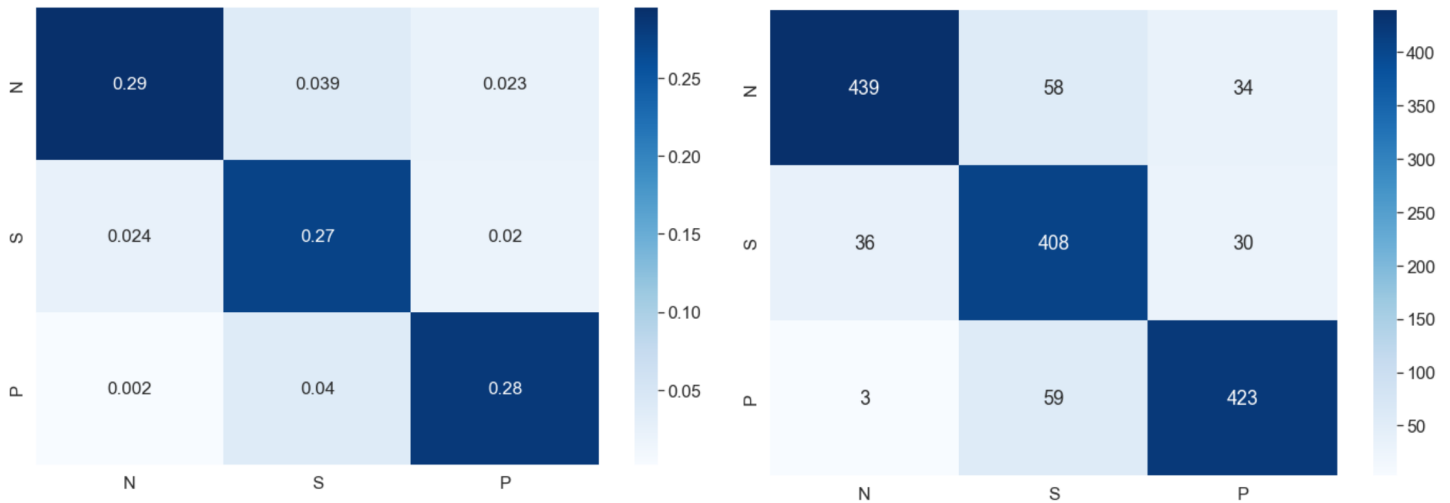|	precision	recall	f1-score	support
1.0	0.92	0.83	0.87	531
2.0	0.78	0.86	0.82	474
3.0	0.87	0.87	0.87	485
accuracy			0.85	1490
macro avg	0.85	0.85	0.85	1490
weighted avg	0.86	0.85	0.85	1490

Figure 8: Random Forest Confusion Matrix

3.2.3 Model 3: Decision Tree

The classifier used to build model 3 is the decision tree classifier that visualizes the decision making process by mapping out different courses of actions and outcome (Myers et al. 2004). Also using the sklearn library, the following parameters were passed into the classifier, "criterion='gini',max_depth=2,random_state=33." With the parsed features with the highest correlations, the model computed an accuracy of 72% and error rate of 28%.

Table 7: Decision Tree Scores

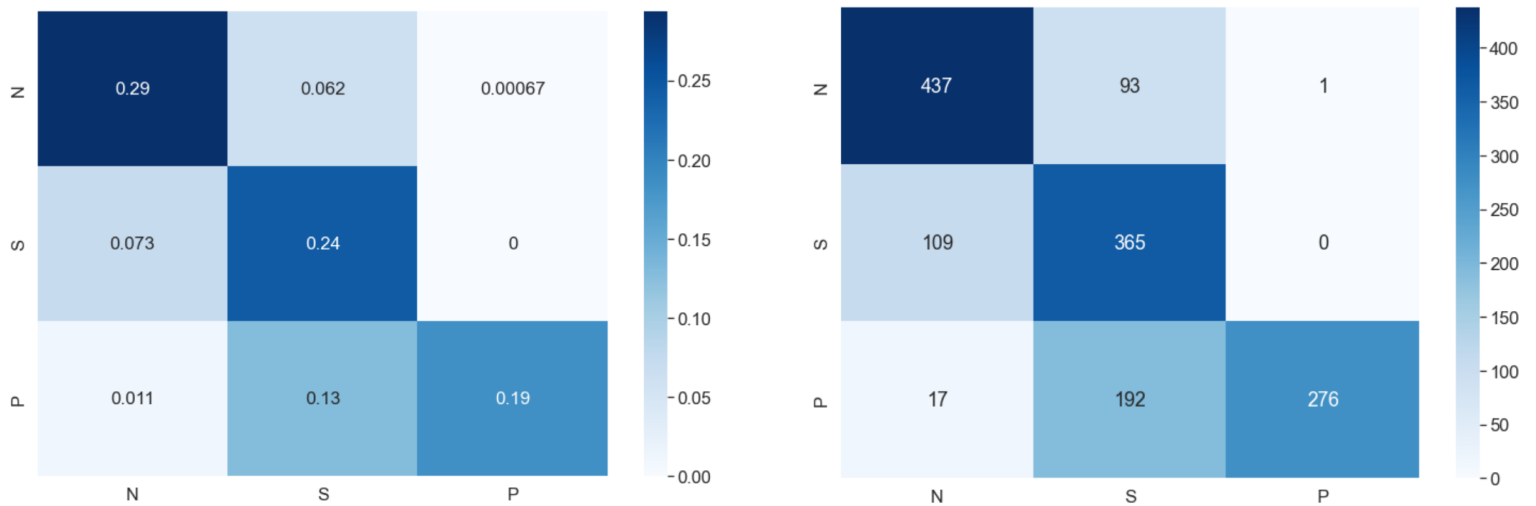|	precision	recall	f1-score	support
1.0	0.78	0.82	0.80	531
2.0	0.56	0.77	0.65	474
3.0	1.00	0.57	0.72	485
accuracy			0.72	1490
macro avg	0.78	0.72	0.72	1490
weighted avg	0.78	0.72	0.73	1490

Figure 9: Decision Tree Confusion Matrix

Results

Out of all of the models, the Random Forest model had the best performance at 85% with the following parameters, "criterion = 'gini',n_estimators =100,max_depth=4,random_state=33." I was able to focus more on this model and improve it by changing the parameters. With the new following parameters, "criterion = 'gini',n_estimators=150,max_depth=12, random_state=33," the model achieved a high score of 96%. In Figure 10, the predicted versus the actual values of this model is represented in graphs. This graph shows that the predicted values from this model

are very similar to the actual values of the data. In Figure 10, the confusion matrix of the improved model with the optimal metrics is represented which explains that the model only falsely predicted 57 times at 4% and correctly predicted 1433 times at the accuracy score of 96%.
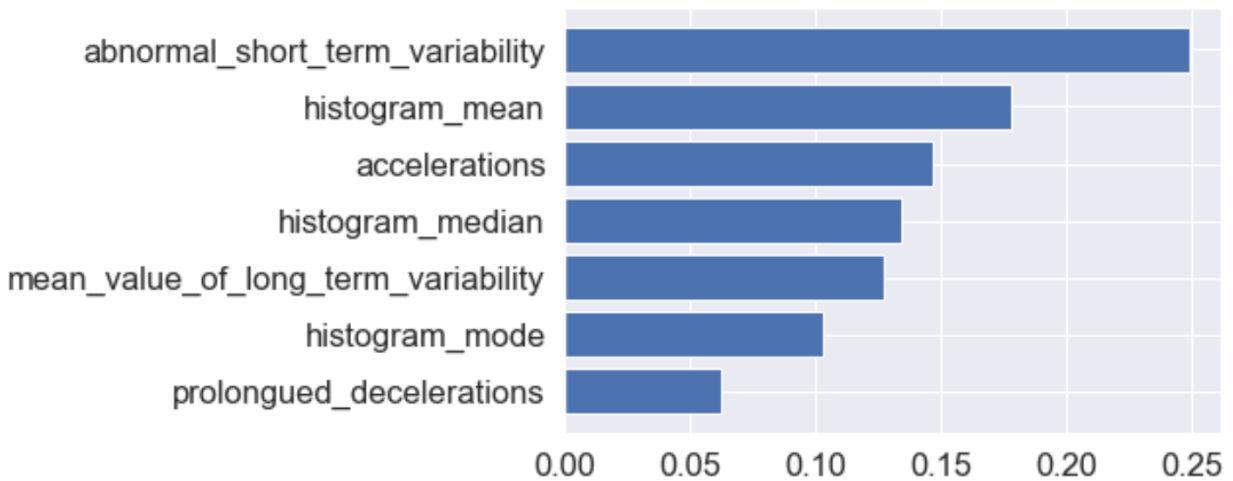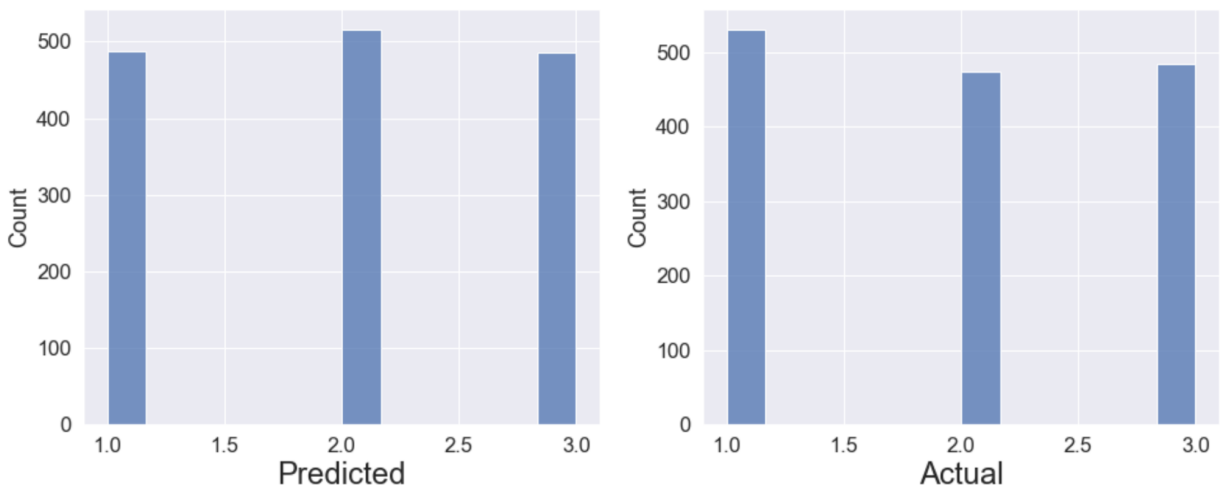

Figure 10: Feature of Importance

Figure 11: Predicted versus Actual Plotted

Table 8: Improved Random Forest Scores

```
              precision   recall  f1-score   support

         1.0       0.99     0.91      0.95       531
         2.0       0.91     0.99      0.95       474
         3.0       0.99     0.99      0.99       485

    accuracy                         0.96      1490
   macro avg       0.96     0.96      0.96      1490
weighted avg       0.96     0.96      0.96      1490
```
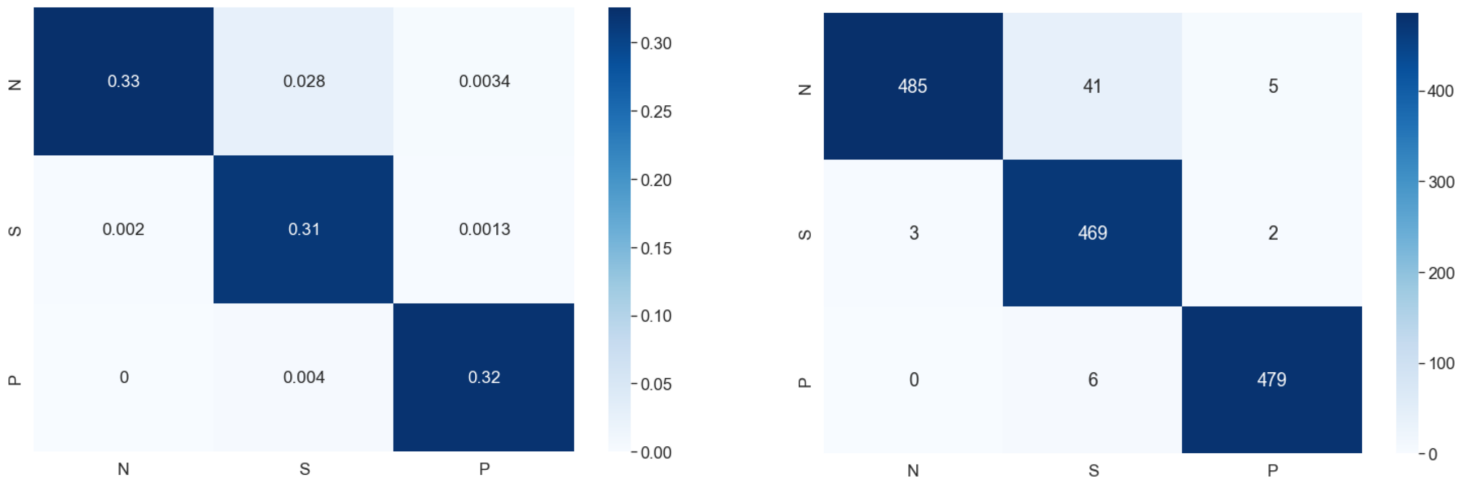

Figure 12: Model 2 Confusion Matrix with Optimal Metrics

Table 9: Accuracy Recap

Classifiers	Accuracy
Linear Regression	81%
Random Forest	85%
Decision Tree	72%
Optimal Random Forest	**96%**

Conclusion

In conclusion the study of the fetal health classification has been reported. Three different models were runned and analyzed using linear regression, random forest and decision tree classifiers. The random forest model performed the best at 85% with seven features that include, "acceleration, prolonged decelerations, abnormal short term variability, mean value of long term variability, histogram mode, mean and median." These features were picked through exploration of the data by looking at the correlations on a heatmap and relationships between predictors and response variables. There were many other correlations with the features but for this classification project, I was specifically looking at the correlations between the features and the response variable of 40% or higher. After getting an accuracy of 85%, I then decided to improve this model and got an even higher score of 96%.

The features that were parsed out based on the heatmap and predictor versus response relationship were all associated with the heart. After looking at the features of importance for the random forest model, abnormal short term variability showed to have the most influence for such a high accuracy. After classifying the data, this tells me that the heart is a key indicator and the recommendation for the target audience is to monitor the heart closely to avoid hypoxia.

Reference

Alam, Mahbubul (12 June 2020). *Dealing with missing data in data science projects*. Towards Data Science.
> https://towardsdatascience.com/dealing-with-missing-data-in-data-science-projects-e8ac7a4efdff

Ayres de Campos et al. (2000) SisPorto 2.0 *A Program for Automated Analysis of Cardiotocograms*. J Matern Fetal Med 5:311-318 (link).
> https://www.kaggle.com/andrewmvd/fetal-health-classification

Aznar, Pablo (2020). Decision Trees: Gini vs Entropy. QuantDare.
> https://quantdare.com/decision-trees-gini-vs-entropy/

Centers for Disease Control and Prevention. 2015. *The Challenge of Fetal Mortality*.
> https://www.cdc.gov/nchs/products/databriefs/db16.htm

Cravit, Rachel (2021). *What is a Decision Tree and How to Make One [Templates + Examples]*. Vengage. https://venngage.com/blog/what-is-a-decision-tree/

DeSisto, Carla; Martin, Joyce; Gregory, Elizabeth; Hoyert, Donna; Valenzuela, Claudia (2021). *Fetal Deaths*. Centers for Disease Control and Prevention.
> https://www.cdc.gov/nchs/data/nvss/webinar/fetal-death-webinar.pdf

Dildy, Gary (2006). *Management of prolonged decelerations*. OBG Management. Pg 30-33.
> https://cdn.mdedge.com/files/s3fs-public/Document/September-2017/1811OBGM_ArticlE2.pdf

Gross, J., & Groß, J. (2003). *Linear regression* (pp. 33-36 Vol. 175). Springer Science & Business Media.
https://books.google.com/books?hl=en&lr=&id=IvAw_1MTASsC&oi=fnd&pg=PR1&dq=linear+regression&ots=GtgL6rawvv&sig=do4Oi2WG5gEDoPjyQE5avnPgvM0#v=onepage&q=linear%20regression&f=false

Elite Data Science (2019). *How to Handle Imbalanced Classes in Machine Learning*. Elite Data
Science. https://elitedatascience.com/imbalanced-classes

Oshiro T.M., Perez P.S., Baranauskas J.A. (2012) *How Many Trees in a Random Forest?*. In:
Perner P. (eds) Machine Learning and Data Mining in Pattern Recognition. MLDM 2012.
Lecture Notes in Computer Science, vol 7376. Springer, Berlin, Heidelberg.
https://doi.org/10.1007/978-3-642-31537-4_13

Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A., & Brown, S. D. (2004). An introduction to
decision tree modeling. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 18(6),
275-285.
https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/cem.873

ProjectPro (2021). *How to deal with imbalance classes with upsampling in Python?* ProjectPro.
https://www.projectpro.io/recipes/deal-with-imbalance-classes-with-upsampling-in-python

Sweha, Amir (1999). *Interpretation of the Electronic Fetal Heart Rate During Labor*. American
Family Physician. https://www.aafp.org/afp/1999/0501/p2487.html

UpToData (2021). *NICHD definitions of FHR characteristics and patterns*. UpToData.
https://www.uptodate.com/contents/image?imageKey=OBGYN%2F65859

Wilson, Debrea (2018). *Fetal Heart Monitoring: What's Normal, What's Not?* Healthline.
https://www.healthline.com/health/pregnancy/abnormal-fetal-heart-tracings