Foetal Health Prediction
(Project Presentation)

Presented by :

Anupam Garg

Bhavya Jain

Shivam Kumar Jha

Subhanshu Bansal

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

Motivation

- Generally during the pregnancy period of a women, it is important to **check the condition of foetus** on a timely basis.
- **Child mortality rates** are often considered as a key indicator of human progress.
- **Cardiotocogram** helps to **monitor the foetus health** and prevents such mortality cases, but there is a risk that **some falsely diagnosed fetus conditions** may lead to some serious problems.
- The main motive behind the study was to **employ some machine learning algorithms** to increase the chances of **correctly classifying the fetus health status**.
- The proposed models can help to **make the doctors' decision more reliable** by providing a simple classification of the fetus health and can help to give a **proper analysis of its condition**.

Literature Review

Foetal Health Prediction Using Classification Techniques[1]

This paper compares different models and predicts foetal health using the processed Cardiotocogram dataset. Cardiotocogram contains distinct signals which are used to record the fetal heart rate conditions and doctors often rely on its results.

Random Forest: which tries to combine multiple decision trees for prediction. it shows less accuracy of **98.33%.**

Support Vector Machines showed an accuracy of **96.54%**, as it divides dataset into various parts as vectors.

Naive Bayes uses probability to predict unknown class and gave accuracy of **97.32%**. **Logistic Regression**, it has accuracy **99.52%** using one/more independent variables with binary outcome.

Literature Review

Foetal Health Classification based on Machine Learning[2]

The author have proposed a model that predicts the foetal health. The dataset used here is the CTG dataset. The model proposed in this paper predicts the foetal health in 3 classes,namely Normal, Suspect and Pathological.

In this paper, the dataset is tested on **12 machine learning models**. In this, K-fold cross validation is used to train the model with K=10. After training and testing the data, the **top four models** were Gradient Boosting Classifier, CatBoost Classifier, Light Gradient Boosting Machine and Extreme Gradient Boosting.

There are two methods, **Blender Model**(soft voting method) and **Stacker Model**. In soft voting method, every classifier's predicted probability for a particular class is weighted according to the individual classifier. The target label having maximum sum is chosen. In Stacker method, learner models are combined with meta models (learn from other learning algorithms).

The **Blender Model** provides with the highest accuracy of 95.9% compared to other models. It also has an AUC of 0.988, recall rate of 0.916 and a precision rate of 0.959.

Dataset Description

The foetal dataset consists of various features as follows:

1. **baseline value** - fetal heart rate baseline(beats per minute)
2. **accelerations** - number of accelerations per second
3. **fetal movement** - number of foetal movements per second
4. **uterine contractions** - no. of times the tightening and shortening of uterine muscles per second
5. **light_decelerations** - no. of times a temporary minor drop in the foetal heart rate per second
6. **severe_decelerations** - it refers to the number of severe decrement in the movements per second
7. **prolongued_decelerations** - these refer to the non-reassuring fetal heart rate characteristics per second
8. **abnormal_short_term_variability**- percentage of time with beat-to-beat variation in foetal heart rate
9. **mean_value_of_short_term_variability-** mean of the short term variability in the FHR
10. **percentage_of_time_with_abnormal_long_term_variability**- refers to the cyclical or rhythmic changes seen in sympathetic nervous system over a minute
11. **mean_value_of_long_term_abnormality-** refers to the mean of the period with cyclical changes in the sympathetic nervous system in a minute

Dataset Description

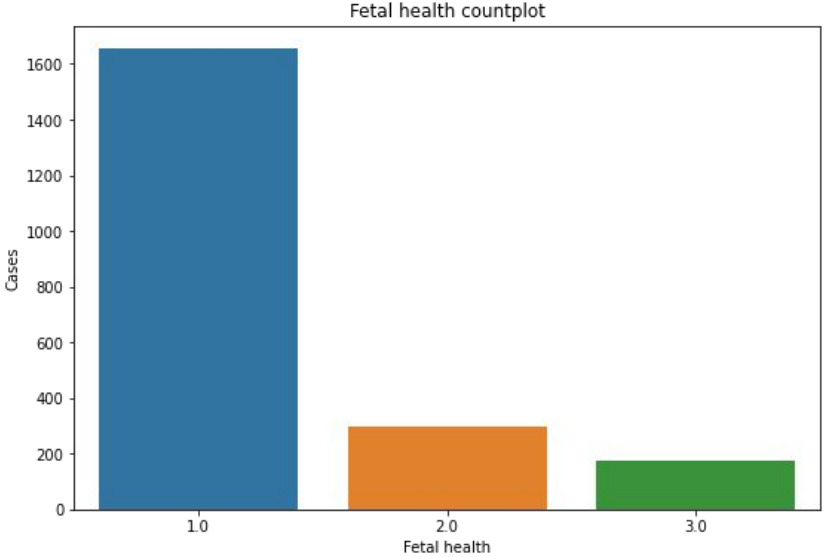
12. **mean_value_of_long_term_viability-** mean value of the long term viability
13. **histogram_width-** width of the fetal heart rate histogram
14. **histogram_min**- maximum(lowest frequency) of the FHR histogram
15. **histogram_max**- maximum(highest frequency) of the FHR histogram
16. **histogram_number_of_peaks**- number of histogram peaks
17. **histogram_number_of_zeroes**- Number of histogram zeros
18. **histogram_mode**- mode of the histogram
19. **histogram_mean**- mean of the histogram
20. **histogram_median**- median of the histogram
21. **histogram_variance**- variance of the histogram
22. **histogram_tendency**- tendency of the histogram
23. **fetal_health**- The features are then classified by obstetricians into 3 classes:

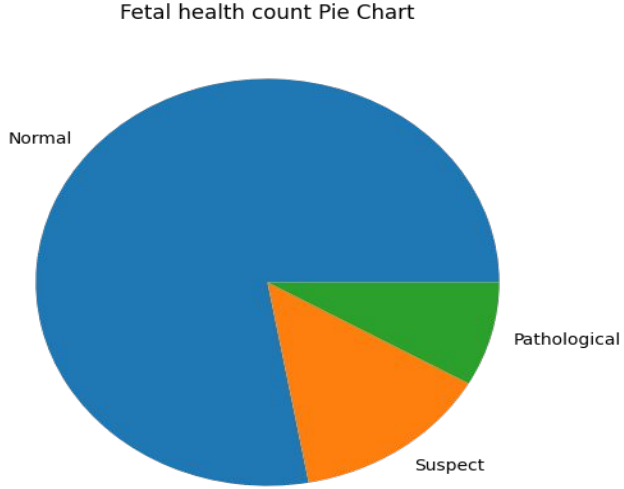
a. Normal (1) b. Suspect(2) c. Pathological(3)

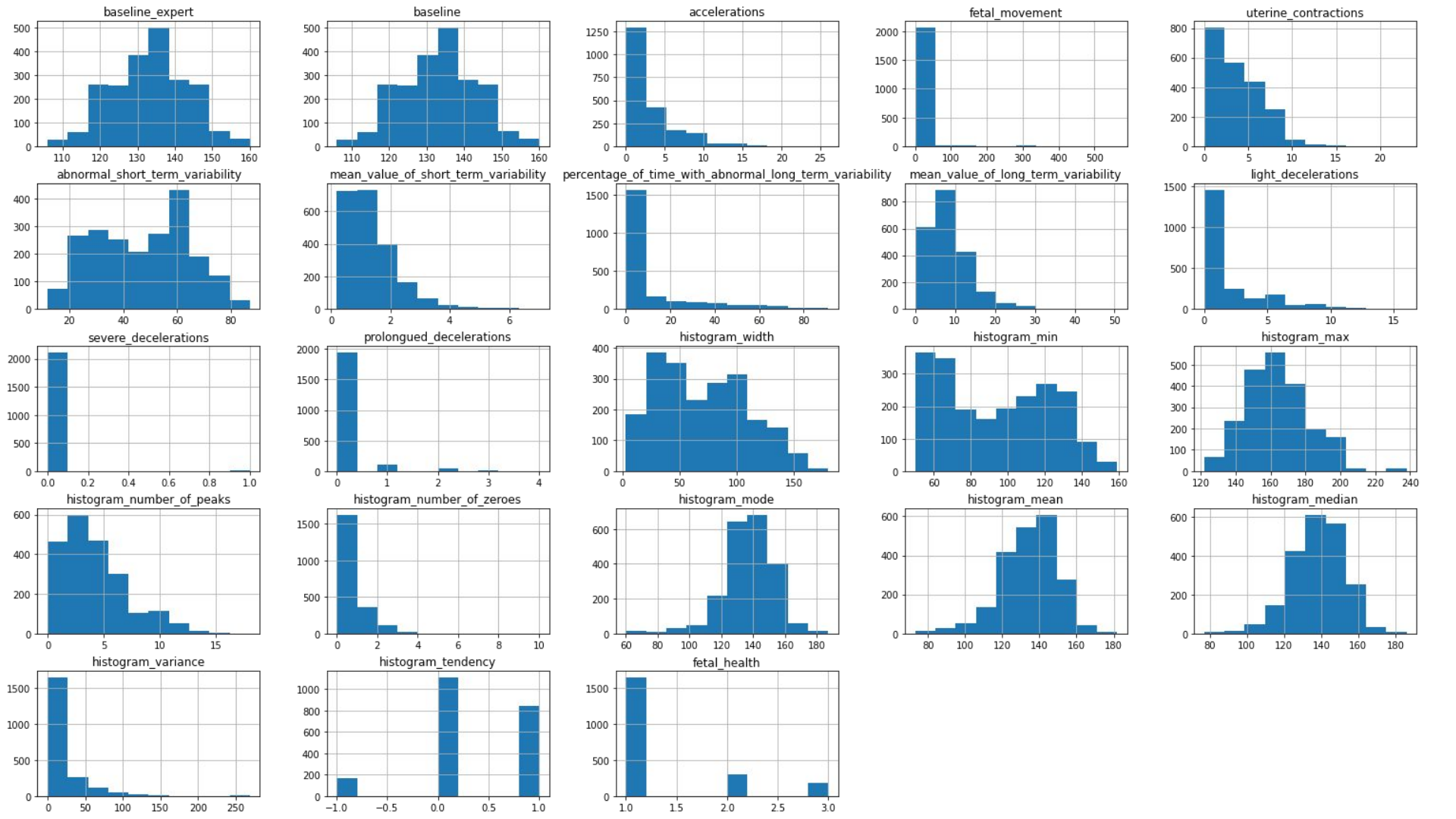
Visualisations and details of dataset

- We have a total of 23 attributes in the dataset, all are continuous and have float dataset.
- On checking for null values, the dataset showed no null values for each of the attribute.

On plotting Histogram and Pie Chart between count of Fetal health cases and labels



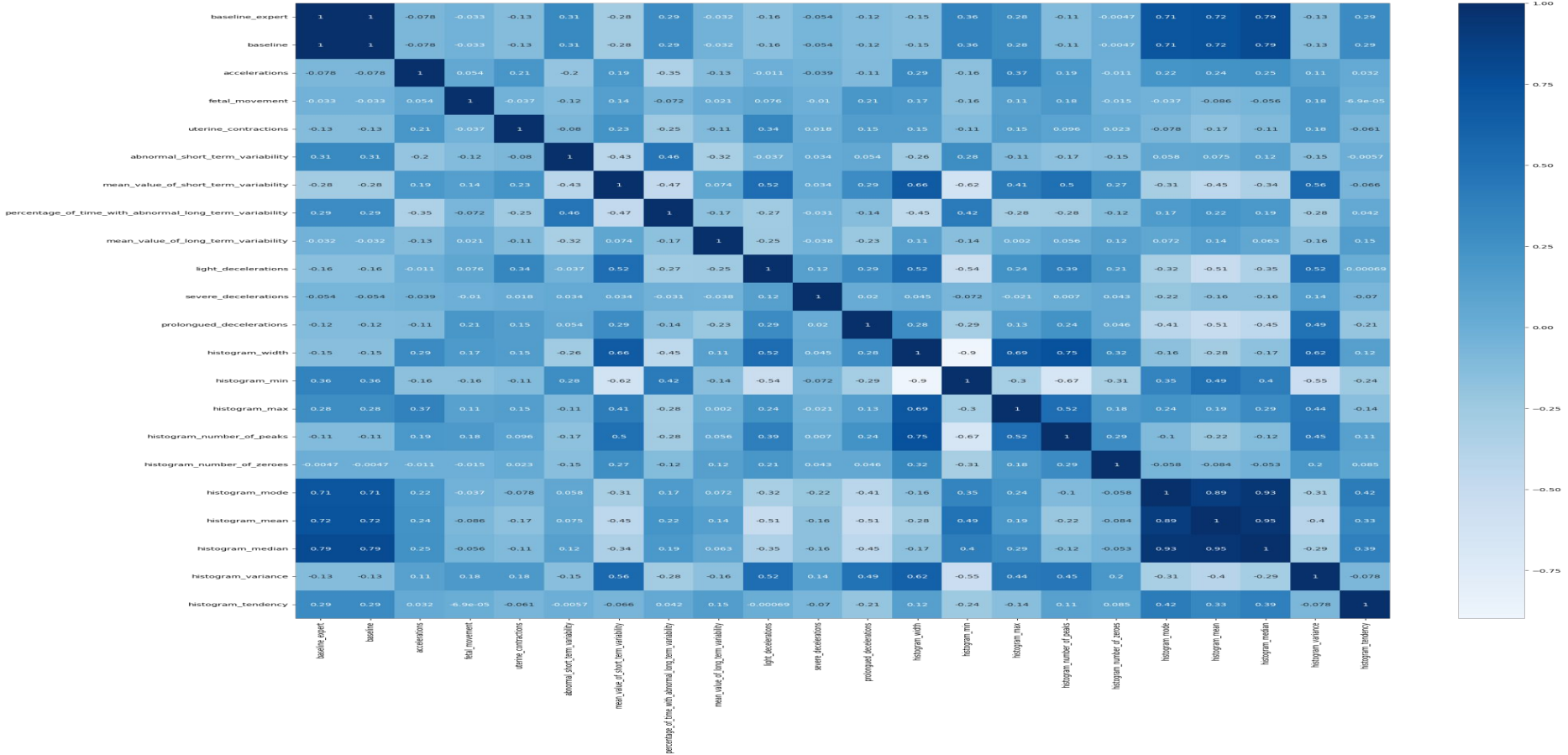



baseline_expert
baseline
accelerations
fetal_movement
uterine_contractions
abnormal_short_term_variability
mean_value_of_short_term_variability
percentage_of_time_with_abnormal_long_term_variability
mean_value_of_long_term_variability
light_decelerations
severe_decelerations
prolongued_decelerations
histogram_width
histogram_min
histogram_max
histogram_number_of_peaks
histogram_number_of_zeroes
histogram_mode
histogram_mean
histogram_median
histogram_variance
histogram_tendency
fetal_health

Data Processing

- Initially, all the irrelevant features from the raw data like filenames, dates etc. were removed.
- As no null values were found, thus there was no need to remove any null values from the dataset.
- Next we tried finding out if there are any two attributes which are highly correlated with each other.
  - To do this we plotted heat maps.
  - Baseline_Expert as it is highly correlated to Baseline.(1)
  - Histogram_Width as it is highly correlated to Histogram_Min (-0.9)
  - Histogram_Median is also highly correlated to Histogram_Mode and Histogram_Mean(0.95,0.93)
  - Thus we dropped these features.

Heatmaps


baseline_expert
baseline
accelerations
fetal_movement
uterine_contractions
abnormal_short_term_variability
mean_value_of_short_term_variability
percentage_of_time_with_abnormal_long_term_variability
mean_value_of_long_term_variability
light_decelerations
severe_decelerations
prolongued_decelerations
histogram_width
histogram_min
histogram_max
histogram_number_of_peaks
histogram_number_of_zeroes
histogram_mode
histogram_mean
histogram_median
histogram_variance
histogram_tendency
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

Data Processing

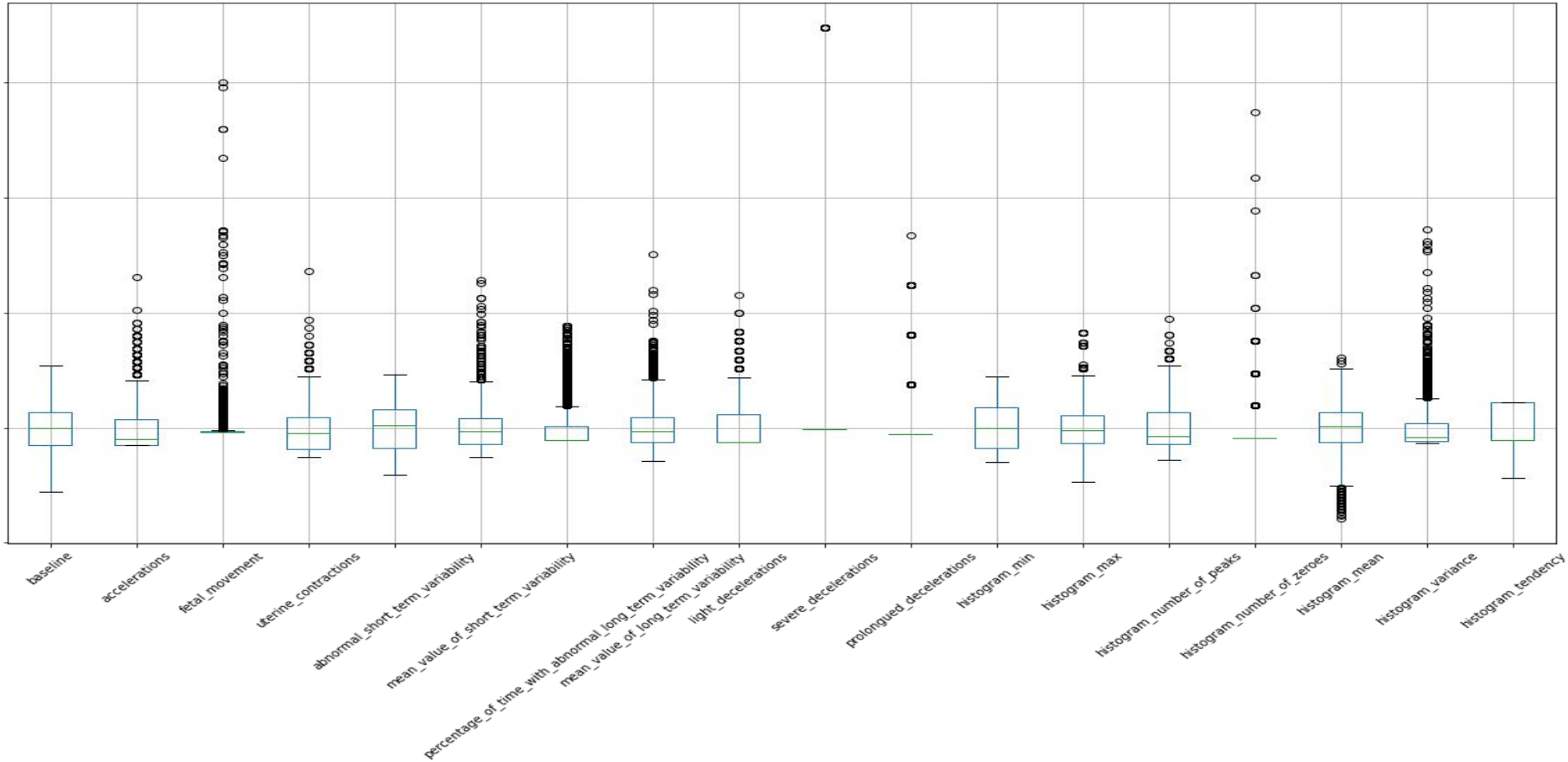
- Then, we scaled our data using standardization.

- **OUTLIER TREATMENT:**

  - Outliers refers to the sample values which varies greatly within the dataset.
  - In order to find the outliers, we made box plots. These box plots clearly indicate which values are varying greatly.
  - Now after finding out the outliers, we made all the cells NULL whose value is more than our threshold(5). We are not removing the rows.
  - Now using the k nearest neighbor technique, we computed these NULL values and filled them with the approximate values.
  - This was done to remove the risk of identifying the correct values belonging to a particular class as outlier.

- **Our Data is Processed Now!**

Outliers in the dataset



Methodology

- **Naive Bayes**: Predicts the class of unknown data sets. Assumes independence among the predictors. Outperforms on highly sophisticated classification methods.
- **Logistic Regression**: Binary outcomes (0/1). Predicts a dependent variable with the help of the given independent variable which are categorical in nature.
- **Random Forests**: Supervised learning algorithm which collects samples from different data sets.Predicts the best solution by combining various decision trees.
- **Adaboost:** Helps in building a model and gives equal weights to all the data points. Reinitialize the weights to each classifier by assigning higher weights to points that are wrongly classified.

Results and Analysis

Our models are evaluated on the basis of accuracy metric. Accuracy measures the overall efficiency of a classifier.

Model	Accuracy
Naive Bayes	80.52%
Random Forests	87.80%
Logistic Regression	86.73%
Adaboost Classifier	87.57%

In our case, **Random Forests** gave the highest accuracy of 87.80%.

Conclusion

- It is important to do exploratory data analysis , data preprocessing and feature engineering to get better results from our models.
- We used k-fold cross validation(k=10) to train our models and used the metric score ‘accuracy’ to evaluate the best one.
- On evaluation, we found random forest to work best with the accuracy score of 87.8%.

Timeline

Our team tried to follow the timeline as per the proposal. Instead of implementing Decision trees due to high computational complexity, we used Adaboost Classifier.

Week 1	Data Collection (including Scraping)
Week 2	Data Pre-Processing
Week 3	Data Visualization
Week 4	Feature Extraction
Week 5	Feature Analysis, Selection, Correlation Matrix
Week 6	Logistic Regression, Naïve Bayes
Week 7	AdaBoost, Random Forest

Future Work

We are expecting to complete the remaining work as per the given timeline below:

Week 8	KNN, Confusion Matrix and Classifier, SVM
Week 9	Analysis and Performance of Models
Week 10	Hyperparameter Tuning
Week 11	Report Writing and Buffer

Team Member Contribution

This is the individual team contribution done till now.

Data Collection	Anupam & Bhavya
Data Pre-processing	Anupam & Bhavya
Data Visualization	Anupam & Bhavya
Feature Analysis	Anupam & Bhavya
Logistic Regression, Naïve Bayes	Shivam & Subhanshu
Random Forest, Decision Trees, AdaBoost	Shivam & Subhanshu
Report Writing	Anupam

Thank You!