Foetal Health Prediction

Project Presentation

Anupam Garg

Bhavya Jain

Shivam Kumar Jha

Subhanshu Bansal

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

Motivation

Fetal Health is an indicator of the fetal wellbeing and regular contact with the womb during the pregnancy period.

Obstetricians often use cardiotocography(CTG) to administer a child's health during the pregnancy period as it helps to diagnose any ailments at initial stages itself.

A complete analysis of CTG analyzed traditional data is time-consuming and unreliable.

A classification model then becomes important as it can monitor the situations faster.

Machine Learning thus becomes useful in such situations as it can address a variety of problems because of its fast development.

The proposed models can help to make the experts decision **more reliable** by providing a simple classification of the fetus health.

Introduction

Medical difficulties and mortality associated with the pregnancy is a matter of concern worldwide, which can affect both the mother and the baby.

Pregnancy problems can include hypertension, diabetes, infection, miscarriage and still birth type problems.

Cardiotocogram is the most widely used technique for recording the fetal heart rate(FHR) and uterine contractions(UC) to evaluate fetal well-being and can detect harmful abnormalities in the newborn baby.

ML tools can help the experts to make informed decisions which can reduce maternal and fetal death rates.

The main aim of the study was to implement different types of ML Algorithms to detect fetal health-related issues relatively faster.

Literature Survey

1. **Prediction of Fetal Health State during Pregnancy[Deressa Tadele Debisa, Kadam Kalyani]**

The paper talks about the importance of fetal health for the mother and the fetus and provides the analysis of efficient machine learning algorithms to accurately predict the outcomes of fetal health.

Classification : It is one of the supervised learning task to predict the health of an unborn baby from the CTG dataset. Artificial Neural Network(ANN), Decision Tree, K-Nearest Neighbor(KNN), Random Forests, Naive Bayes(NB), Support Vector Machine(SVM) and Cross Validation are some of the data-mining algorithms used for the data classification.

Literature Survey

1. **Artificial Neural Network(ANN):** It composes of number of input layers, hidden layers and output layers w.r.t weight which transfer signals between neurons.
2. **Support Vector Machines(SVM)**: It is a classifier which performs classification tasks by constructing the hyper-planes in multidimensional space which separates case of different class labels.
3. **K-Nearest Neighbor Algorithms(KNN):** It is an algorithm which is based on the distance between the items, which uses Euclidean distance among pair of data in N-dimensional space.
4. **Random Forest Algorithm(RF)**: It tries to fit a number of decision-tree classifiers using bagging methods.
5. **Cross Validation**: The method provides upoptedness to choose a specific part of the training and the testing set.

Literature Survey

The accurate precision response was calculated by dividing the dataset into 10 equal size. ANN is applied with high accuracy for very large and complex dataset. ANN method is preferred for the prediction of the hypertension disease. Its high computation cost and high learning rate makes it inefficient. SVM classification method classifies the data into binary or multi-classes by searching the best hyperplane which classify the data under proper category. A linear SVM was chosen as the classifier to look at the relationship between features and adverse outcomes.

Predicting the risk mitigation based on the model and for particular patient trend was the limitation of the paper.

Literature Survey

2. Foetal Health Classification based on Machine Learning[Li Jiaming, Lui Xiaoxiang, IEEE 2021]

The author have proposed a model that predicts the foetal health. 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**.

The **top four models** were Gradient Boosting Classifier, CatBoost Classifier, Light Gradient Boosting Machine and Extreme Gradient Boosting.

There are two methods, **Blender Model** and **Stacker Model**.

In Blender Model method, every classifier's predicted probability for a particular class is weighted according to the individual classifier.In Stacker method, learner models are combined with meta models that 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

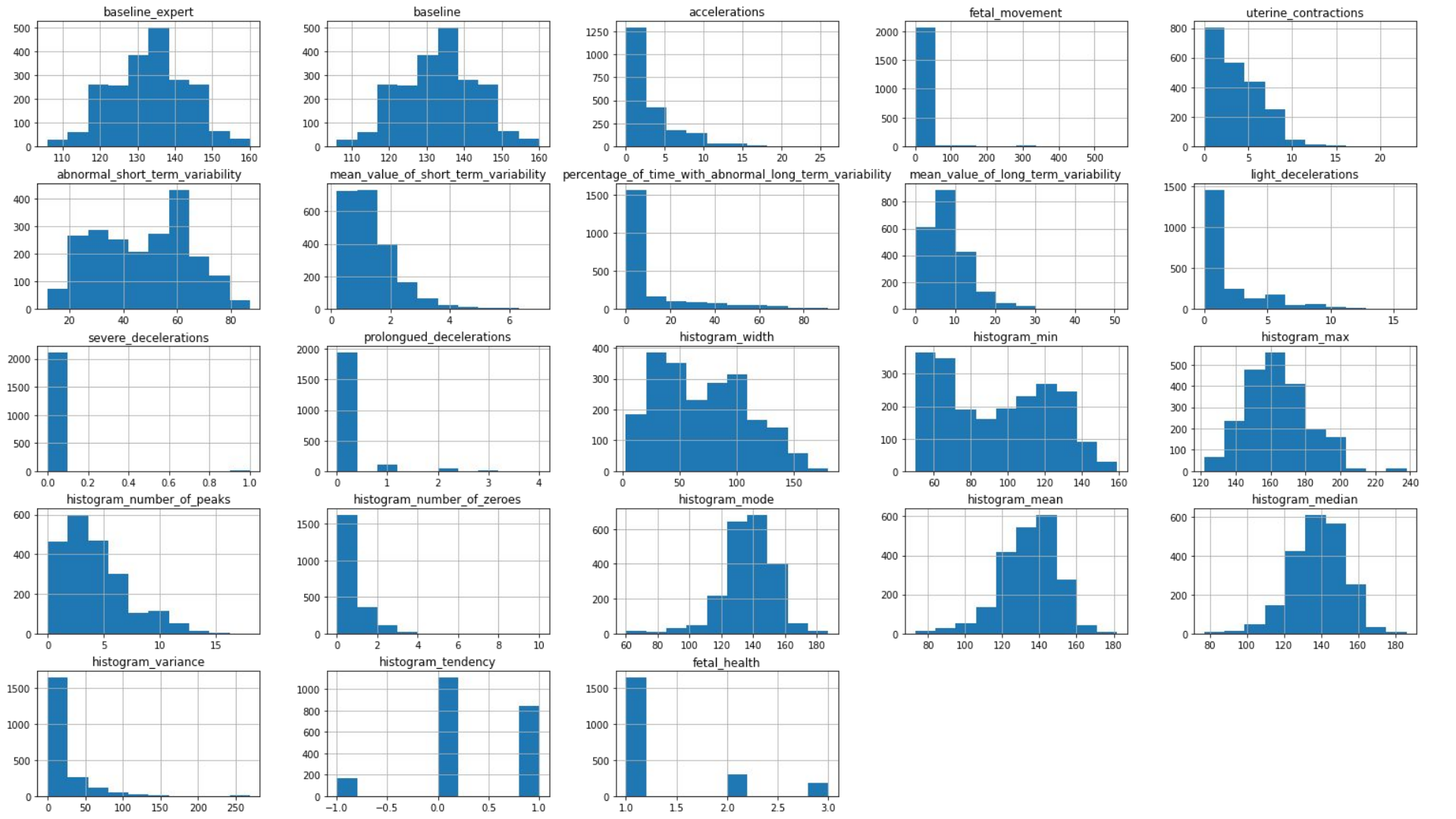
The foetal dataset consists of various features as follows:

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:

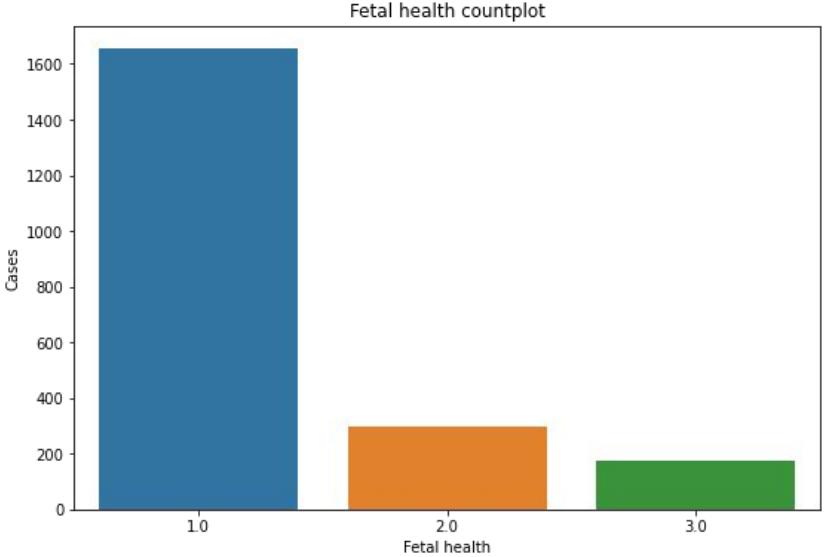
a. Normal (1) b. Suspect(2) c. Pathological(3)

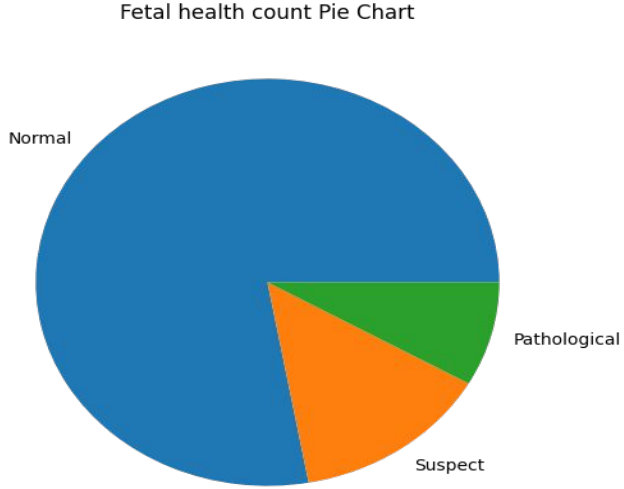

baseline_expert
baseline
accelerations
fetal_movement
uterine_contractions
abnormal_short_term_variability
mean_value_of_short_term_variability
percentage_of_time_with_abnormal_long_term_variability
mean_value_of_long_term_variability
light_decelerations
severe_decelerations
prolongued_decelerations
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

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





Data Processing

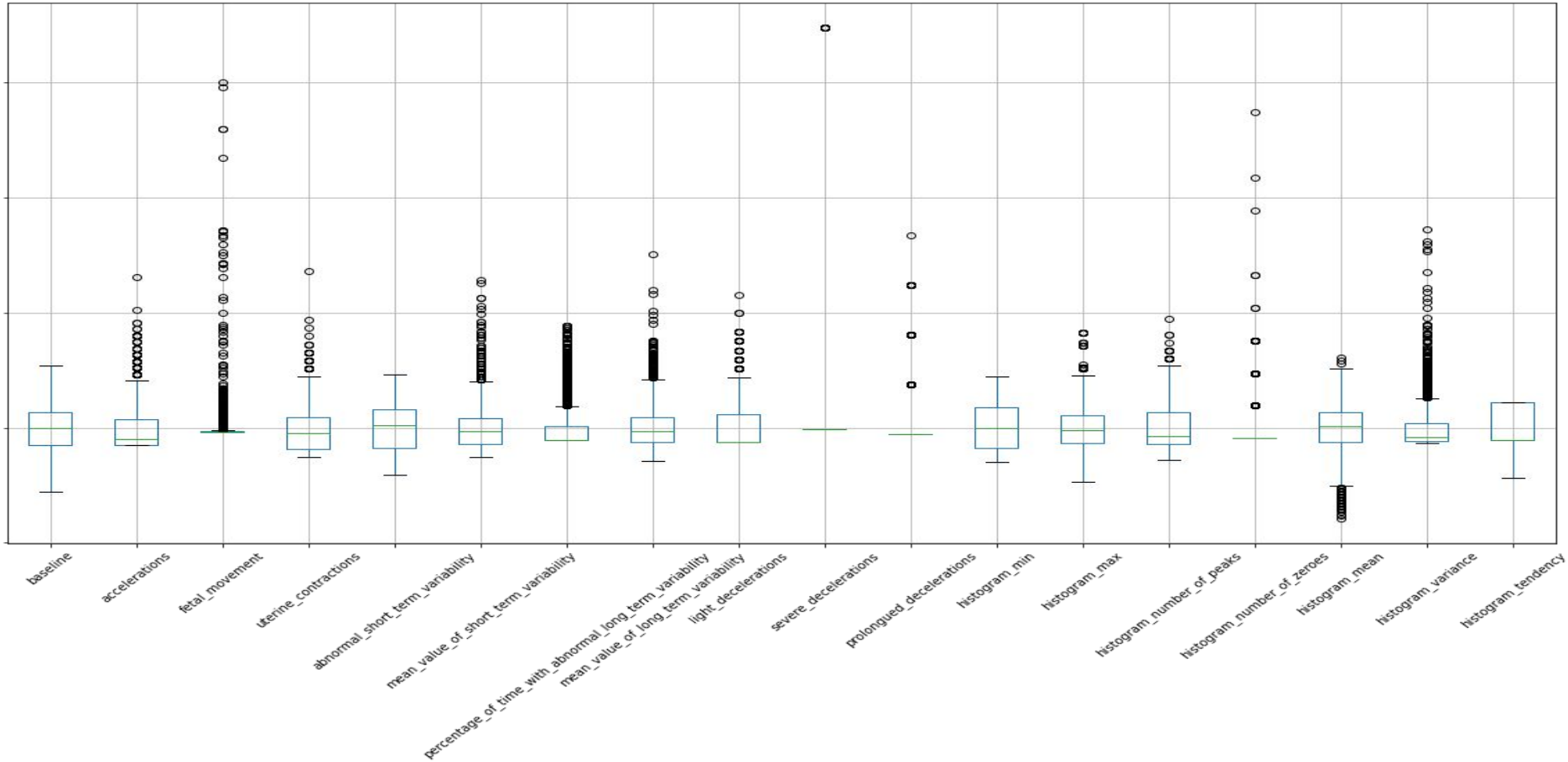
- Then, we scaled our data using standardization.

- **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 and Model details

- **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.

Methodology and Model details

- **Artificial Neural Network**: Artificial Neural Networks are made up of layers and layers of connected input units and output units called neurons. Signals are modified at the receiving synapse and the weighted inputs are summed at the processing element.he activation function is a transfer function that is used to get the desired output for the problem designed.In our case , classify the input into different sets.

- **K-Nearest neighbour**:It is one of the machine learning algorithm to classify the input in different set.It classifies the new data points based on the similarity measure of the earlier stored data points

- **Support Vector Machine**: support vector machine algorithm is to find a hyperplane in an n-dimensional space that distinctly classifies the data points

Results and Analysis

Our models are evaluated on the basis of f1-score metric. F1-score is more better than accuracy as it imposes a penalty on the incorrectly classified samples.

a. **Logistic Regression:** 0.7466

b. **Naive Bayes:** 0.6670

c. **Random Forests:** 0.8105

d. **Adaboost Classifier:** 0.7577

e. **Artificial Neural Network(ANN):** 0.8120

f. **Support Vector Machine(SVM):** 0.7255

g. **K-Nearest Neighbors(KNN):** 0.7047

Conclusion

- Health complications during pregnancy time constitute a significant challenge confronted across the globe.

- This work assesses the influence of various factors measured through CTG to predict the health state of the fetus through algorithms like SVM, RF, MLP, and K-NN.

- We used k-fold cross validation(k = 10) to train our models and used the metric **f1-score** to evaluate the best one.

- On evaluation, we found ANN to work best with the f1-score value of about 0.8120.

Team Members Contribution

This is the individual team contribution for the proposed project.

Data Collection	Anupam & Bhavya
Data Pre-processing	Anupam & Bhavya
Data Visualization	Anupam & Bhavya
Feature Analysis	Anupam & Bhavya
Methodologies and Model training	Shivam & Subhanshu
Random Forest, Decision Trees, AdaBoost	Shivam & Subhanshu
Report Writing	Anupam, Subhanshu and Shivam

Thank You!