FETAL DISTRESS CLASSIFICATION USING CARDIOTOCOGRAPHY

S.Santhiya, G.Kowreesh, G.V.Chitraleka, P.Boobalaragavan
UG Scholar, UG Scholar, UG Scholar, UG Scholar,
Department of Software Systems,
Coimbatore Institute of Technology, Coimbatore, India.

ABSTRACT

Children are the future of the world. So, this work wants to be sure that they are born without any complications and disabilities. But here there are many new born infants who suffered from complicated disorders like fetal distress, which is the complication which arises during delivery. This condition is basically caused because of insufficient oxygen supply and lack of other vital factors inside their mother's womb. This distress will lead to some dangerous disorders like neurological malfunctions and sometimes it can also lead to fatal. Cardiotocography is an effective, efficient and widely used method to analyze the fetal heart beat rate (FHR) and uterine contraction of the mother and so on. This cardiotocography plays an important role in analyzing various fetal distress conditions and their possibilities. So, let us built a machine learning model which analyses the results and observations made during the cardiotocography and predicts the possibilities of the fetal distress to be occurred.

KEYWORDS

Cardiotocography, Fetal distress, Random forest, Decision Trees, Fetal Heart Rate

I. INTRODUCTION

Fetal distress is a known emergency condition which occurs during the labor and delivery. It is a complication in which a baby experiences Hypoxia (occurs when inadequate supply of oxygen to brain) and Birth asphyxia (oxygen deprivation at tissue level of body). This can even changes the baby's heart rate, decreases movement of the fetus, and also decreases the meconium level in the amniotic fluid and in some serious cases it can also causes hypoxic-ischemic encephalopathy (HIE) and permanent injury to the fetus. So, Signs of fetal distress should not be taken lethargically. And Medical professionals must be aware of it before to take immediate actions and manage the situation and at the same time it is critical that medical professionals promptly recognize and address these signs. So it is necessary for he/she to be careful especially when assessing maternal and fetal health during the cases of high-risk pregnancies. So, there is a need for the continuous monitoring of fetal health throughout the pregnancy using cardiotocography (also known as Electronic Fetal monitor (EFM), a commonly used method to monitor the Fetal heart rate (FHR) and uterine contraction) and others like examining the fetal movement in womb, amniotic fluid level. This paper performs some analyses by using the observed data. That is the need here and as we all know that need is the seed for invention, this work is to develop a fetal distress classifier by cardiotocography using the machine learning algorithm. Section 2 is Data analysis it contains exploration of data and its key features. Section 3 is Algorithm and technique it contains working principle of random forest algorithm and workflow of this model. Section 4 is Methodology it contains data preprocessing (selection of highly important columns) along with step by step process for constructing this model.

II. DATA ANALYSIS 2 DATA EXPLORATION

This CTG (Cardiotocography) dataset consist of 2126 records of various fetus's cardiotocography reports. More accurately there are 40 columns presents to determine this data. This data has been classified by three expert obstetricians and they were labelled accordingly. Here the classifications are based on respect to their morphological patterns and their fetal states. There are 10 morphological pattern's classes namely A,B,C… SUSP and three fetal state classes namely N,S, and P that is Normal, Suspect and Pathologic respectively.

Table 2.1: Columns present in the CTG dataset

File Name	File name of cardiotocography examination	NMAX	No. of histogram peaks
Date	Date of examination of CTG	NZEROS	No. of histogram zeros
b	start instant	MODE	Mode of histogram
E	end instant	MEAN	mean of histogram
LBE	baseline value (medical expert)	MEDIAN	Median of histogram
LB	baseline value (SisPorto)	VARIANCE	Variance of histogram
AC	accelerations (SisPorto)	TENDENCY	Tendency of histogram: -1=left asymmetric; 0=symmetric; 1=right asymmetric
FM	fetal movement (SisPorto)	A	calm sleep
UC	uterine contractions (SisPorto)	B	REM sleep
ASTV	% of time with the abnormal short term variability (SisPorto)	C	calm vigilance
MSTV	mean of short term variability (SisPorto)	D	active vigilance
ALTV	% of time with the long term abnormal variability (SisPorto)	SH	shift pattern [class A or Suspect with the shifts}
MLTV	The mean of the long term variability (SisPorto)	AD	Accelerative or decelerative pattern [stress situation]
DL	decelerations – light	DE	decelerative pattern - vagal stimulation
DS	decelerations – severe	LD	decelerative pattern - large
DP	decelerations - prolonged	FS	pathological state - flat-sinusoidal pattern
DR	decelerations - repetitive	SUSP	pattern of suspect
WIDTH	width of histogram	CLASS	Class code (1 to 10) for classes A to SUSP
MIN	lower frequency of histogram	NSP	Normal=1; Suspect=2; Pathologic=3
MAX	higher frequency. of histogram		

III. ALGORITHMS AND TECHNIQUES

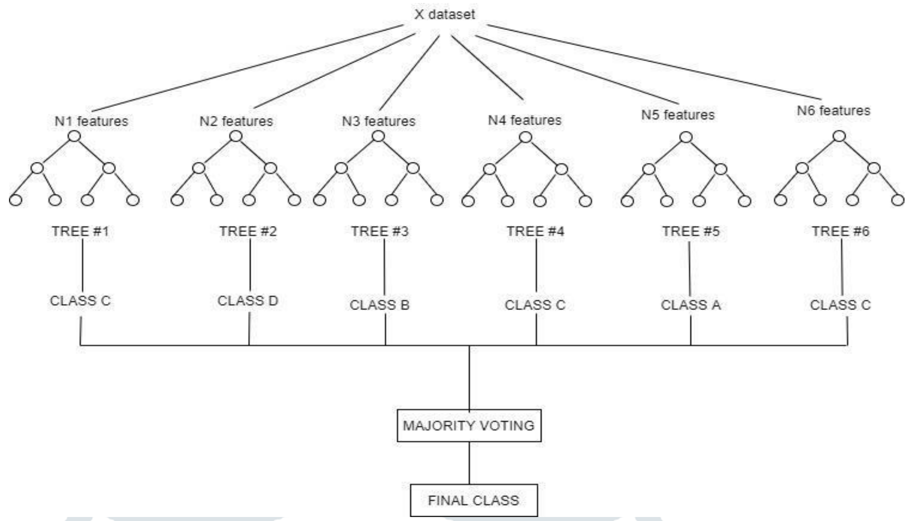
3.1 RANDOM FOREST ALGORITHM

The one of the most famous supervised learning algorithms is Random forest, which is used for both classification and also regression. Each forest is made up of trees and as we all know that if there are more trees means that is a more robust forest. Similarly, here also random forest algorithm creates many decision trees on the given data samples and then get prediction result from each of the created decision tree and then it finally selects the best solution among the predicted solution by voting. It is known as ensemble method which is known for the solution which is better than the solution of a unit decision tree because it tends to drastic reduction of over-fitting by averaging the result. The random forest algorithm has several advantages like they work very well for large range of data sample rather than a single decision tree. It also has less variance, very flexible and possess very high accuracy even when there is missing in large portion of data. And it also doesn't require scaling since it can produce a good accuracy even without scaling.

The working principle of Random Forest Algorithm includes,

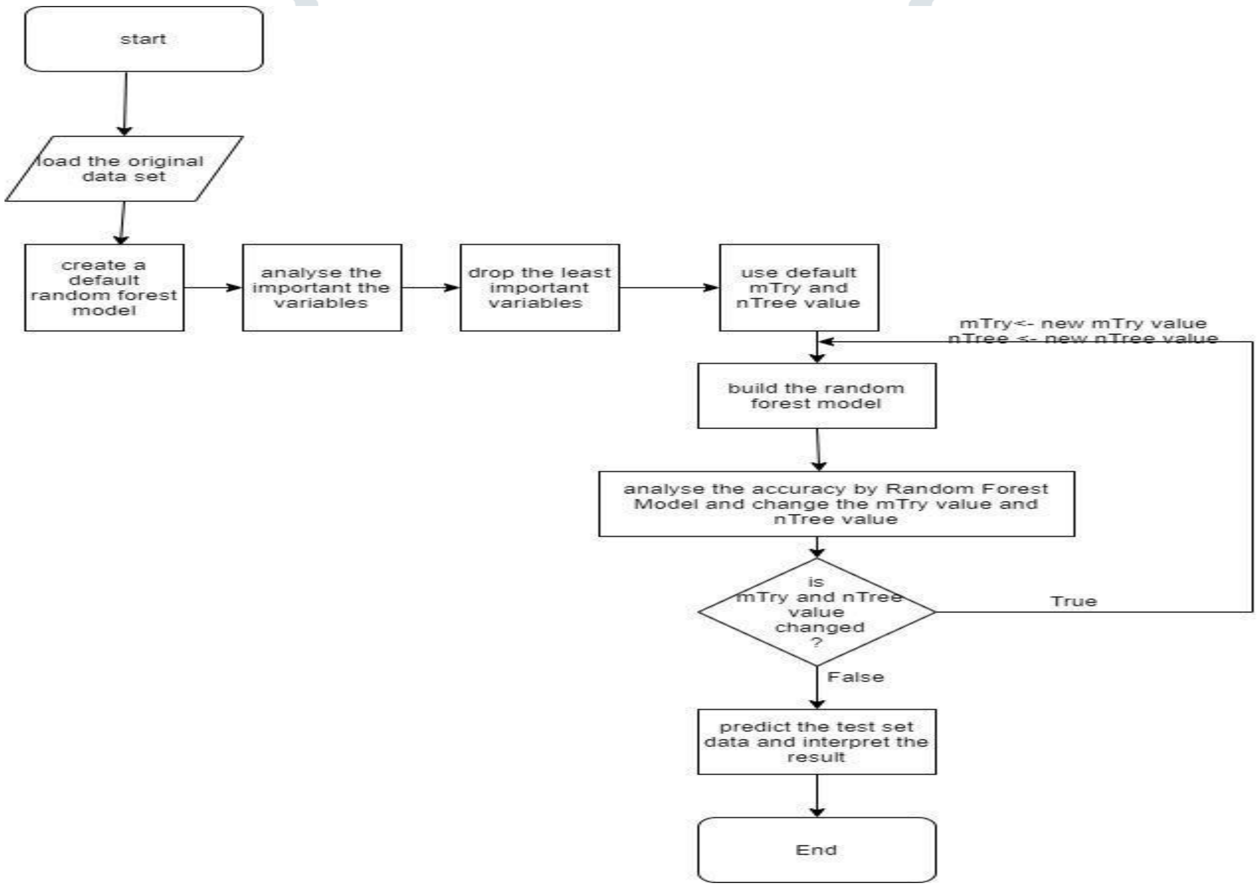
i. **Step 1** − Selection the random samples from the given dataset.
ii. **Step 2** – Then the algorithm will construct some number of decision tree for every sample. Then it predicts the result from the created decision trees.
iii. **Step 3** – The voting is performed by the predicted results.
iv. **Step 4** – The most voted result in the prediction is taken as the final result.

Figure 3.1: Random forest algorithm



3.2 WORKFLOW USED FOR THE CONSTRUCTION OF MODEL

Figure 3.2: Workflow of the model



IV. METHODOLOGY

4.1 DATA PREPROCESSING

The data preprocessing is done by using the following steps. They are,

i. The data set is first loaded and the data is analyzed

ii. Then the null values in the dataset are identified and removed by using **dropna()** function.

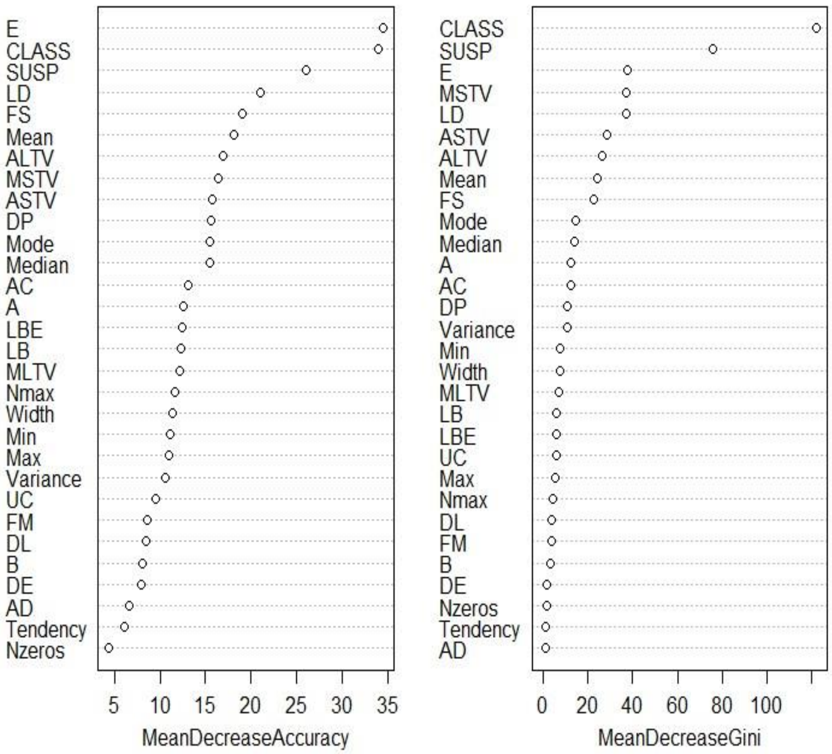
iii. The data types of the available columns are identified and scaling is performed.

iv. Convert the NSP column from continuous value to categorical variable by using **as.factor()** function v.

Dropping the least important columns:

- **Step-1** => Construct a Random Forest model with default values for mTry and nTree.
- **Step-2** => Construct the importance plot for the newly created model to identify the least important features. The below are the plot of the important features.

Figure 4.1: Extracted important columns



- **Step-3** => Extract the top 21 features and NSP column from the original dataset to make the model more optimized.

  The Extracted features are, "LB", "AC", "FM", "UC", "DL", "DS", "DP", "ASTV", "MSTV", "ALTV", "MLTV" , "Width", "Min", "Max", "Nmax", "Nzeros", "Mode", "Mean", "Median", "Variance", "Tendency", "NSP"

So here we reduced the dimensionality of the data set and from now onwards we are going to process and predict by using only the extracted columns.

4.2 CONSTRUCTION OF RANDOM FOREST MODEL

Firstly, Constructing a random forest model with default parameter as,
Number of trees = 500 and Number of variables tried at each split = 4.
Here the Out Of Bag (OOB) estimate of error rate is observed to be 5.64%

The confusion matrix of this model is

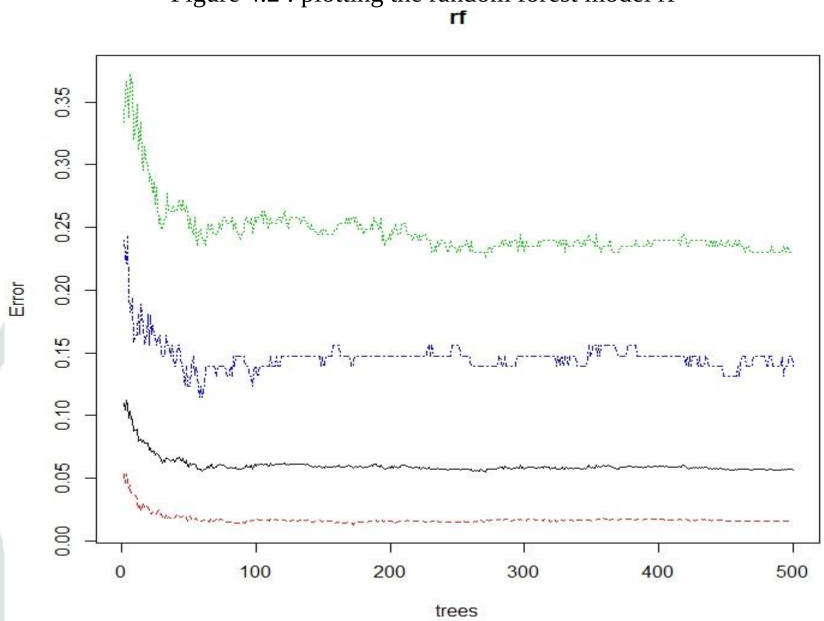
Table 4.2: Confusion matrix of random forest model with default values for mTry and nTree

Prediction/Reference	1	2	3	CLASS ERROR
1	1140	16	4	0.01724138
2	46	164	3	0.23004695
3	7	8	107	0.12295082

4.2.1 MODEL OPTIMISATION

- Firstly, Determine the perfect optimum value for the number of decision trees to be used. To obtain that lets plot the above created random forest model rf and find the optimal value using the plot.

Figure 4.2 : plotting the random forest model rf



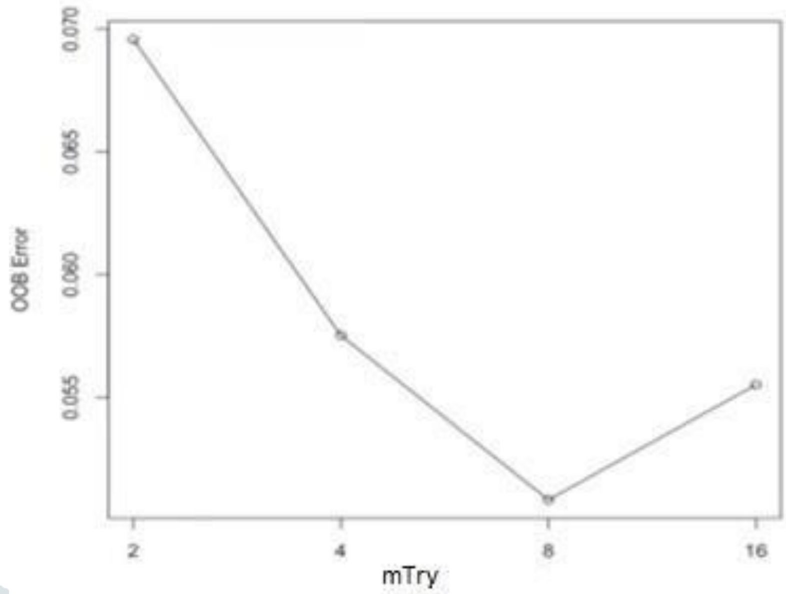
- In the above plot, the green, blue and red represents the classes pathologic, suspect and Normal respectively. And the black line represents the normalization of error. And from the above plot it can be concluded that the 300 decision trees will serve us the good model.

- Secondly, Determine the optimum value for the number of variables tried at each split by tuning the mTry variable as and store it in variable 't' follows:

```
t <- tuneRF (train [, -22], train [,22],
                    stepFactor = 0.51, plot = TRUE,
 ntreeTry = 300, trace = TRUE, improve = 0.05)
```

The resultant plot is :

Figure 4.3: Relationship between the OOB Error and mTry



From the above plot we can clearly observe that the mTry value of 8 gives the least Out of Bag error when compared to other values. So, lets finalize the mTry value as 8.

4.2.2 RECONSTRUCTION OF THE MODEL

- Now we construct the new random forest model by using the optimized values of the number of variables tried at each split (mTry) and Number of trees (nTree) values.
- That is ,

```
rf <- randomForest (NSP~.,data=train,
            ntree    =    300,
mtry = 8,
            importance    =    TRUE,
proximity = TRUE)
```

○ Then the resultant model has the Out of Bag Error of 5.08% and the confusion matrix of this model is Table 4.3: Confusion matrix of optimized random forest model

	1	2	3	CLASS ERROR
1	1137	16	7	0.01982759
2	41	171	1	0.19718310
3	6	5	111	0.09016393

4.2.3 PREDICTING THE TEST DATA

Let us use the optimized random forest model for predicting the result for the test dataset and the resultant gave us Accuracy of 93.5%

And the sensitivity of the classes 1,2 and 3 are 0.9677, 0.75610 and 0.90741 respectively. The below shown table is the confusion matrix of the test data on the optimized model:

Table 4.4: Confusion matrix of optimized random forest model of the test dataset

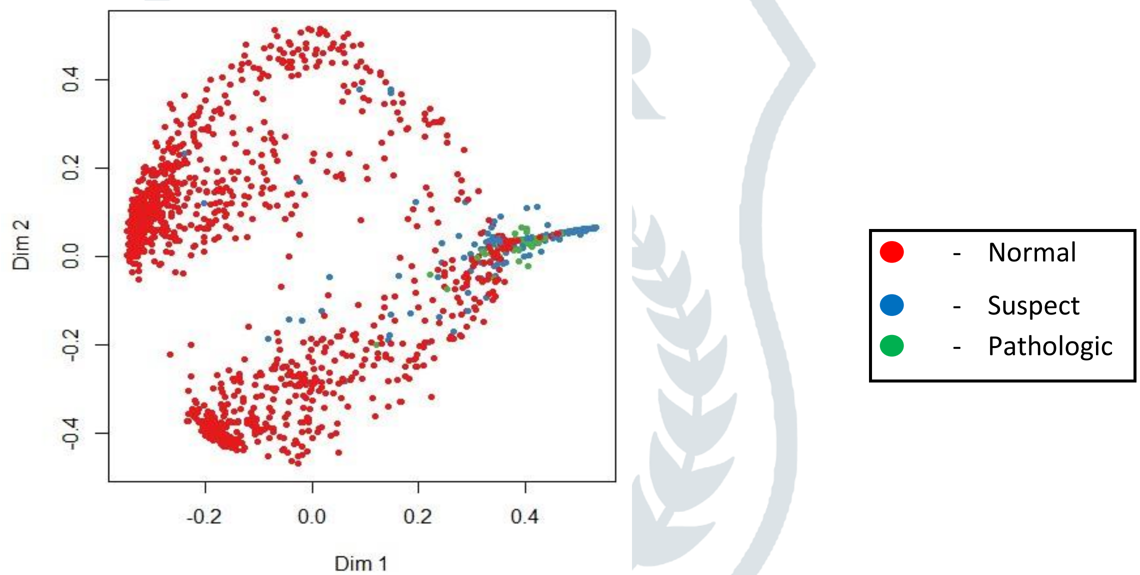
Prediction/Reference	1	2	3
1	479	15	3
2	14	62	2
3	2	5	49

And it also provides the balance accuracy of 0.9177, 0.86348 and 0.94764 for the classes namely 1,2 and 3 resp ectively.

V. RESULT

By constructing the random forest model by using the default parameters for mTry and nTree for fetal distress classification using cardiotocography we obtained the accuracy of 94.46% for training dataset. After tuning the default parameters we got an accuracy of 94.92% for training dataset and accuracy of 93.5% in test dataset. Below shown scatter plot is the representation of prediction made by the optimized model on test dataset.

Figure 5: Scatter plot of the predicted data points



VI. CONCLUSION

The great advancements in the medical area are mostly based on the computer based decision support. This advancements requires a major contribution of the technologies like machine learning for preventing the human errors and to improve the quality of the medical treatments. Sometimes these solutions can fail while dealing with different problems. Therefore we are in need of wide search area to predict the solution and this is very vital factor while it is coming to medical decision. So here we uses the results of the prior analysis of several CTG in determination of the state of the fetal distress. Hence Here we made an computerized approach on the cardiotocography data using a machine learning algorithm called random forest to classify the fetal distress into three major classes namely Normal, Suspect and Pathologic and we obtained the optimal results.

REFERENCE

[1] https://www.abclawcenters.com/frequently-asked-questions/what-are-some-signs-that-my-baby-is-in-distress/#:~:text=Fetal%20distress%20is%20an%20emergency,amniotic%20fluid%2C%20among%20other%20signs.

[2]https://www.researchgate.net/publication/334042156_Fetal_Distress_Classification_based_on_Cardiotocography_using_Machine_learning

[3]https://www.researchgate.net/publication/272670242_Analysis_of_Cardiotocogram_Data_for_Fetal_Distress_Determination_by_Decision_Tree_Based_Adaptive_Boosting_Approach

ABOUT AUTHORS

S .SANTHIYA, III Year MSc Software Systems, Department of Software Systems, Coimbatore Institute of Technology, email-santhiyasundarraj05@gmail.com

G .KOWREESH, III Year MSc Software Systems, Department of Software Systems, Coimbatore Institute of Technology, email-kowreesh3810@gmail.com

G. V. CHITRALEKA, III Year MSc Software Systems, Department of Software Systems, Coimbatore Institute of Technology, email-chitraleka892@gmail.com

P. BOOBALARAGAVAN, III Year MSc Software Systems, Department of Software Systems, Coimbatore Institute of Technology, email-boobal520@gmail.com