CSE / ECE 343: MACHINE LEARNING INTERIM SEMESTER PROJECT REPORT

Title: Foetal Health Prediction

Anupam Garg
anupam20555@iiitd.ac.in

Bhavya Jain
bhavya20428@iiitd.ac.in

Shivam Kumar Jha
shivam20332@iiitd.ac.in

Subhanshu Bansal
subhanshu20135@iiitd.ac.in

Abstract

During the child birth, it is important to check the condition of foetus on a timely basis. Child mortality rates are important for the human progress. The reduction of child mortality is shown in several of the UN sustainable goals and is a crucial indicator of human progress. Cardiotocogram(CTG) helps to monitor the foetus health and prevents such mortality cases, but there is a risk associated with it that some falsely diagnosed conditions can lead to some serious problems. The proposed machine learning models can help to make the doctors' decision more reliable by providing a simple classification of the foetus health and can help to give a proper analysis of its condition.

Introduction

It is important to get the information about the foetus during the pregnancy period but it is tough to get the accurate signals. CTG contains distinct signals and is mainly used for foetal heart rate recordings. Obstetricians mainly rely on the information from the instruments to give any conclusion. Latest trend observed by doctors is that very high variations are observed in foetal heart rate patterns. But there is a risk that falsely diagnosed foetal pain may lead to major problems, hence the main motive of the research is to employ machine learning algorithms to classify the methods. The prediction algorithm may perform really well in this case and can help monitor the results and give a proper analysis than the doctor could get by his or someone's observation.

Literature Review

Foetal Health Prediction Using Classification Techniques [link]

This paper talks about the importance of accurate diagnosis of the foetus. CTG contains distinct signals which are used to record the foetal heart rate conditions and doctors often rely on its results. Some latest trends by doctors shows that there are very high fluctuations in foetal heart rate patterns, which associates a risk of a false positive prediction and becomes unreliable in some cases.

In this paper, various methodologies have been discussed like the Random Forest, which tries to collect all the dataset and combines multiple decision trees for prediction. But, because it is quite complex, it shows less accuracy of **98.33%**. On the other hand, Support Vector Machines showed an accuracy of **96.54%**, as it divides dataset into various parts as vectors and separates them using a hyperplane. Naive Bayes uses probability to predict unknown class and gave accuracy of **97.32%** while for Logistic Regression, it has accuracy **99.52%** using one/more independent variables with binary outcome. Logistic Regression performed better out of all the four classifiers. There is scope of better results if more training data could be sampled. Also, the comparisons could be made on the accuracy values predicted using models.

Foetal Health Classification based on Machine Learning [link]

In this paper, the dataset used here is the CTG dataset. Prenatal monitoring of CTG consists of signals like foetal heart rate (FHR) and uterine contractions (UC). The model proposed in this paper predicts the foetal health in 3 classes, Normal, Suspect and Pathological.
In this paper, the dataset is tested on **12 machine learning models**. After training and testing the data, the **top four models** were Gradient Boosting Classifier, CAT Boost Classifier, Light Gradient Boosting Machine and Extreme Gradient Boosting. The author integrated these models using 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).
Thus, finally on comparing these 2 combined models, the **Blender Model** has the highest accuracy of 95.9% compared to the rest classifiers.

Dataset Processing

1.Data Preparation

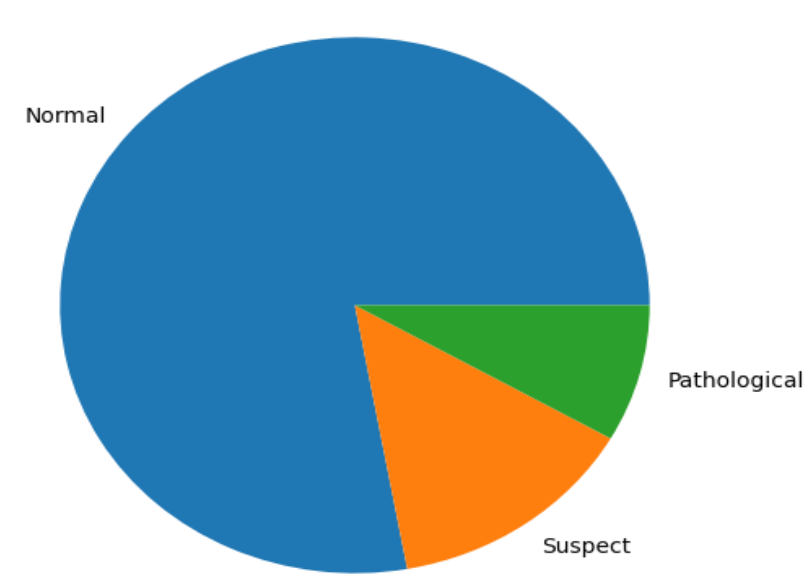
The foetal dataset predicts foetal health using multiple characteristics such as baseline value, accelerations, foetal movement, uterine contraction, light decelerations, etc. For data visualization, the filtering approach is used to choose a subset of the original train data.

2. Data Pre-Processing

A dataset is made up of patterns or entities. A set of characteristics that characterize data items captures an item's essential features. We first remove the null values to improve the accuracy of the dataset. Using boxplots, we detected the outliers and replacing them with null values, and further computed them using K-nearest neighbour technique.

3. Feature Engineering

Feature engineering plays an essential role as everything from the data to the output of the same depends on the feature engineering performed. A correlation matrix of all the features to their significance is used in getting the relation of each feature or attribute and the other features present in the database. Highly correlated features can be removed to reduce the redundant features in the dataset.



Model Methodologies

A. Logistic Regression

In Logistic Regression there are only binary outcomes (0/1). It is used to predict a dependent variable with the help of the given independent variable which are Categorical in nature.

B. Naïve Bayes

The Naïve Bayes algorithm is used to predict the class of unknown data sets. It considers an assumption of independence among the predictors. It is also used to outperform on highly sophisticated classification methods.

C. Random Forests

Random Forest Algorithm is a supervised learning algorithm which collects samples from different data sets and predicts the best solution by combining various decision trees. The main disadvantage of using RFs is the slow computation in prediction and complexity.

D. Ada-boost Classifier

It is a boosting algorithm which helps in building a model and gives equal weights to all the data points. It reinitializes the weights to each classifier by assigning higher weights to points that are wrongly classified. Now all the points which have higher weights are given more importance in the next iteration. It will keep improving training models till less errors are received.

Result/Analysis

As we have implemented all the algorithms using K-fold cross validation having K=10.

We found that the best results are obtained in Random Forest model with accuracy of 87.80% and with AdaBoost having an accuracy of 87.57%. Using Naïve Bayes Algorithm, the accuracy was obtained as 80.52%. On implementing L.R(Logistic Regression), we have the model accuracy as 86.73%.

Conclusion

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

Individual Tasks

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

Timeline

Week 1: Data Collection (including Scraping)
Week 2-3: Pre-Processing and Data Visualization
Week 4: Feature Extraction
Week 5: Feature Analysis, Selection, Correlation, Heatmaps
Week 6: Logistic Regression, Naïve Bayes
Week 7: AdaBoost Random Forest
Week 8: K-Nearest Neighbours, Confusion Matrices & Classifiers, Support Vector Machine
Week 9: Analysis & Performance of models
Week 10: Hyperparameter Tuning, Check for Underfitting & Overfitting
Week 11: Report Writing
Week 12: Buffer

Remaining Work

To work on machine learning models like SVM, K - Nearest Neighbours, Confusion Matrices & Classifiers, Hyperparameter Tuning, Check for Overfitting & Underfitting.

References

[1] Foetal Health Prediction using Classification Techniques (link)
[2] Foetal Health State Monitoring Using Decision Tree Classifier from Cardiotocography Measurements. (link)
[3] Foetal health status prediction based on maternal clinical history using machine learning techniques (link)
[4] Use of Machine Learning Algorithms for Prediction of Foetal Risk using Cardiotocographic Data (link)
[5] Cardiotocography Analysis for Foetal State Classification Using Machine Learning Algorithms (link)