CSE / ECE 343: MACHINE LEARNING FINAL 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 childbirth, it is important to check the condition of the fetus on a timely basis. Child mortality rates are important for 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 fetus' 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 fetus' health and can help to give a proper analysis of its condition. [GITHUB]

Introduction

It is important to get the information about the fetus 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 fetal heart rate patterns. But there is a risk that falsely diagnosed fetal 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 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 fetal heart rate conditions Latest trends by doctors show that there are very high fluctuations in fetal 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 gives 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 fetal heart rate (FHR) and uterine contractions (UC). The model proposed in this paper predicts the fetal 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 the 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 the 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

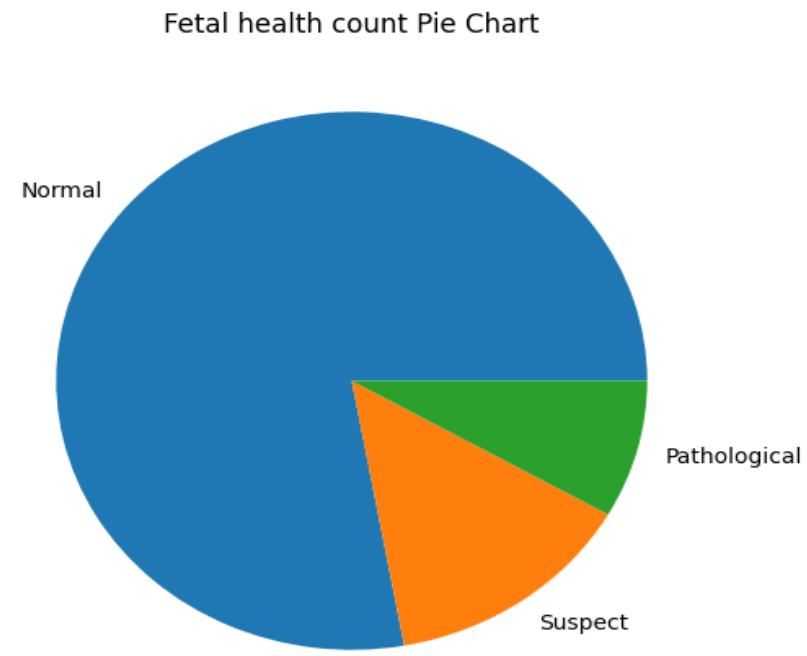
The foetal dataset predicts foetal health using multiple characteristics such as baseline value, accelerations, foetal movement, uterine contraction, light decelerations, etc. For data visualization, the filtering approach is used to choose a subset of the original train data.

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 replaced 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 is 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.

E. K-Nearest Neighbors

KNN is a non-parametric and lazy learning algorithm. Non-parametric i.e., there is no assumption for underlying data distribution. In other words, the model structure is determined from the dataset. It is based on real world datasets which do not follow mathematical theoretical assumptions.

F. Artificial Neural Network

It is a supervised learning model based and inspired from human brain cells also called neurons. It is a collection of functioning computing systems called neurons consisting of various inputs and one output to various other neurons. The last neuron of this collection helps in getting the desired output result.

G. Support Vector Machine

It is a supervised machine learning model and maps training examples to points in space so as to maximise the width of the gap between the two categories. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

H. Cross Validation

CV is the most important method to examine the efficiency of the model by performing a crossover check in successive rounds. The method provides a way to choose a specific part of the training and testing set. In k-fold CV the dataset is divided into k partitions, and repeat k times that each k teams reserved for validation and the other $k-1$ has been used for building the model.

Result/Analysis

As we have implemented all the algorithms using K-fold cross validation having K=10. We used f_score with average macro to compare our models.

We found that the best results are obtained in the ANN model with an f1-score of 81.20%. Next best result was obtained with Random forest with an f_score of 81%. While using AdaBoost, we got the f_score as 75.77%. Using Naïve Bayes Algorithm, the f_score was obtained as 66.71%. On implementing L.R(Logistic Regression), we have the model accuracy as 74.66%.

Furthermore, we found that with SVM the f_score with “linear” kernel 72.55%, along with “rbf” kernel 73.56% and with “poly” 74.23%. After this we used KNN model and the f_score came out as 70.47%. Thus, we get the best result from ANN as it has the ability to learn and implement non linear complex relationships.

Conclusion

1. It is important to do exploratory data analysis, data pre-processing and feature engineering to get better results from our models.
2. We used k-fold cross validation(k=10) to train our models and used the metric score ‘accuracy’ to evaluate the best one.
3. With ANN, we found the best f_score of 81.20% and it is more than the previous ones.
4. These models can be very useful in predicting the health of the fetus and help in taking necessary steps.

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)