An Approachable Introduction to Data Science Case Study: FETAL HEALTH CLASSIFICATION USING CARDIOTOCOGRAPHIC (CTG) DATA

Adewale Adebogun Odunsi, *Registration Number: PG 2110481*
School of Computer Science and Electronic Engineering,
University of Essex, Colchester CO4 3SQ, UK

Abstract— According to World Health Organization, maternal mortality is still unacceptably high, reiterating that about 295,000 women died during and following pregnancy and childbirth in 2017 alone. The report also says that the vast majority of these deaths (94 percent)occurred in low-resource settings, and that most could have been prevented by early detection of complications during pregnancy and childbirth. Clearly one reason why such needless deaths occur in low-resource settings is due to little or no access to medical equipment like a cardiotocography and overworked health care professionals. Thankfully a lot of the developed economies donate equipment to developing ones however there still remains the challenge of insufficient number of obstetricians. Technology has made it possible to use cardiotocographic information presented in a data table that has been vetted and categorized by experts for training a supervised machine learning algorithm, thereby, making such algorithm an 'expert' in predicting the health of a fetus as as obstetrician would. An expert system like this would reduce the workload on such healthcare professionals as it would allow them focus more on critical cases. Added to this but more than this, low resources areas of the would where there are shortfalls in healthcare professionals can leverage on this technology to reduce maternal mortality. In this project a comma separated file dataset that has been categorized by obstetricians is used to train 5 models and the predictions analyzed for correctness. Feature extraction using correlation matrix and cross validation was employed in the design of the models. Ensemble learning model with an average performance score of 93 percent was chosen over a single high performing model like Random forest which has an average performance score of 94 percent. This is because, although Random forest performed marginally higher than the Ensemble model on F1 score, precision, recall and accuracy; the performance was not significant enough to outweigh the benefits of voting among the 5 models.

Index Terms—WHO, Maternal Mortality, MMR, Cardiotocography, CTG, Ultrasound, Fetus, Machine Learning, Ensemble Learning, Random Forest, Decision Tree, Support Vector Machine, Ridge Classifier, K Nearest Neighbor

1 INTRODUCTION

FIVE facts according to World Health Organisation (WHO) are that;

(1) Every day in 2017, approximately 810 women died from preventable causes related to pregnancy and childbirth.[1]

(2) Between 2000 and 2017, the maternal mortality ratio (MMR, number of maternal deaths per 100,000 live births) dropped by about 38 percent worldwide.[1]

(3) 94 percent of all maternal deaths occur in low and lower middle-income countries.[1]

(4) Young adolescents (ages 10-14) face a higher risk of complications and death as a result of pregnancy than other women.[1]

(5) Skilled care before, during and after childbirth can save the lives of women and newborns.[1]

WHO report further said maternal mortality is still unacceptably high reiterating that about 295 000 women died during and following pregnancy and childbirth in 2017. The vast majority of these deaths (94 percent) occurred in low-resource settings, and that most could have been prevented by early detection of complications during pregnancy[1]

Cardiotocography (CTG) is a technology leveraged by healthcare professionals in observing the heartbeat of a fetus as well as contractions of the uterus of a pregnant mother. The device employed is an electro-mechanical machine called a cardiotocograph. With this device, well-being and early detection of fetal distress is made possible with each regular appointment of the expectant mother to the hospital.[2]

The device works by using sound waves at an ultra-high frequency (ultrasound) detectable only by special machines. Ultrasound travels easily through fluid and soft tissues bouncing off internal organs. A sensitive receiver records tiny changes in the sound's pitch and direction. Graphic soft wares measures and displays these signature waves as real-time pictures on a monitor and as information recorded on a paper strip (cartograph).[2]

External cardiotocography is use for both continuous and intermittent monitoring. Two transducers placed on the

- *Wale Odunsi is with School of Computer Science and Electronic Eng., University of Essex, Colchester CO4 3SQ, UK*
E-mail: ao21969@essex.ac.uk

Submitted to CSEE, University of Essex on April 14, 2022 for assessment by: Dr. Haider Raza (email: h.raza@essex.ac.uk).

mother's abdomen and above the fetal heart, monitors heart rate. A third transducer placed at the fundus of the uterus measures the uterine muscle activity, frequency of contractions and the fetal heart. However, internal cardiotocography on the other hand uses an electronic transducer, directly connected to the fetus. A wired scalp electrode from a monitor, is attached through the mother's cervical opening to the fetal scalp[3]

The world population keeps increasing putting pressure on health care workers who are increasingly becoming fewer relative to the population. Is there a possibility of augmenting the shortfall in healthcare specialists with a computer? Could machine learning be employed to douse the pressure on specialist? Could a computer be use to foretell fetal healthcare conditions in areas where there are no specialist or at least refer only serious cases to specialist? This project builds a machine learning models as assistants to health care specialist using fetal health data, available from previously analysed CTG data and refers only suspect and pathological conditions to a specialist for re-examination[3].

2 Literature review

Computers have been around for a long time. Computers contain hardware and software that allows the machine do its job. Hardwares are the physical part of the machine we can see, touch and kick around. Parts like monitors, keyboard, mouse, printers etc make up the Input/Output(I/O) devices. Others like Hard disc, RAM, flash drive make up the memory devices. The processor is the component that does execution using I/O and memory.[4]
Software is the algorithms built with a programming language which the processor follows to execute a task. Traditionally, the programs take input like data, process them and give output as information. Once a program is written, it continues to receive input and output in the format prescribed and can only be altered by a re-writing of the program code.[5]

2.1 Machine Learning Approach

Another approach is to develop algorithms that solve problems by learning from previously available data. This concept is what is referred to as machine learning. Three types of machine learning approach exist.

First is a supervised learning in which the algorithm is ”supervised” by a label or ground truth. In this approach, the machine learns from a sample in which the output is known apriori and is able to generalise for unseen data.

The unsupervised approach learns by clustering samples based on learned patterns in the data while the third, re-enforcement learning is achieved by reward and punishment as the algorithm explores an environment. Computer games and robots make use of the third in their operation.[6]
This project makes used of supervised learning technique in training algorithms to learn patterns in a data-sets giving them the ability to make predictions when presented with new future data-sets. Elaborately, the algorithms will be trained on an existing data-set which has been vetted and categorized by obstetricians and the algorithm becomes a model which like the medical expert, can look at a new patient presentation and predict the condition of the fetus as ”Normal”, ”Suspect” or ”Pathological”.

3 Methodology

Several machine learning algorithms have been developed over the years. In supervised learning for example, algorithms like Support Vector Machines, Discriminant Analysis, Naive Bayes, K-nearest neighbor, Decision Trees and Neural Networks have been use to solve regression and classification problems[7]. The task in this project is a classification problem, one that involves predicting a class or category. As such, the project will look at algorithms that address classification.

3.1 The Dataset

The dataset is a comma separated value (CSV) file consisting of 2126 rows of records and 22 columns of features. A CSV file can be viewed by any spreadsheet or text editor program[8]. This particular CSV file consist of samples drawn from several cardiotocogram examinations done by technologist and classified into three classes by expert obstetricians. These experts are the right persons in decisions and their categorisation is taken as the ground truth. According to the experts, the classification classes are
1.0 Normal
2.0 Suspect
3.0 Pathological

The data-set is loaded into the python program using a data science library called pandas. Pandas project created by Wes McKinney, is a flexible, fast and powerful open source tool for data analysis and manipulation, designed for use within Python programming language.

3.2 Exploratory Data Analysis [EDA]

The data-set was checked for missing data and outliers. This involves, confirming that the data-set has 2126 rolls and 22 columns, inspecting the datatypes for appropriateness, because, any datatype of value ”object” can create a manipulation challenge for the algorithm. The datatype were all of type float64 and thankfully there were no Null/NaN values in the database. If Null/NaN values are present, replacing such values with a measure of central tendency like the mean, mode or median of the values in the column, or simply removing that record would addressed this.

Next, a matrix of Correlation Coefficients was generated. This is a matrices of numbers that indicates relationships among data features as well as the ground truth (label). A strong positive correlation is one in which there is a strong direct relationship between the features in such a way that, as one feature increases, the other increases as well. An inverse relationship is true for negatively correlated features. The project drilled further to see which features are strongly correlated with the ground truth (label) and to prune out those features with very weak correlation with the ground truth.

CORRELATION MATRIX:

.	fetal-health
1. accelerations	-0.364066
2. histogram-mode	-0.250412
3. histogram-mean	-0.226985
4. mean-value-of-long-term-variability	-0.226797
5. histogram-median	-0.205033
6. uterine-contractions	-0.204894
7. histogram-tendency	-0.131976
8. mean-value-of-short-term-variability	-0.103382
9. histogram-width	-0.068789
10.histogram-max	-0.045265
11.histogram-number-of-peaks	-0.023666
12.histogram-number-of-zeroes	-0.016682
13.light-decelerations	0.058870
14.histogram-min	0.063175
15.fetal-movement	0.088010
16.severe-decelerations	0.131934
17.baseline value	0.148151
18.histogram-variance	0.206630
19.perc.-of-time-with-abnorm.-long-term-variab.	0.426146
20.abnormal-short-term-variability	0.471191
21.prolongued-decelerations	0.484859
22.fetal-health	1.000000

Table 1: Correlation of features with ground truth

From the correlation coefficients matrix, It is reasonable to drop features from the table whose correlation coefficients are between -0.075 and +0.075 because clearly, these have very weak linear correlation with the ground truth (fetal-health) and as such can be prune out. Therefore the following features were dropped:

FEATURES	CORR.with LABEL
1. histogram-width	-0.068789
2. histogram-max	-0.045265
3. histogram-number-of-peaks	-0.023666
4. histogram-number-of-zeroes	-0.016682
5. light-decelerations	0.058870
6. histogram min	0.063175

Table 2: Pruned features due to weak correlation

Two most positively correlated features relative to label are prolongued-decelerations(0.484859) and abnormal-short-term-variability(0.471191)
Two most negatively correlated features are accelerations(-0.364066) and histogram-mode(-0.250412)

Categorical plots show that
1. Low fetal movement and low prolongued decelerations are indicative of a normal fetal health while moderate to high fetal movement and a high prolongued decelerations are indicative of pathological fetal health.
2. Very low accelerations with significantly high fetal movement are indicative of pathological fetal health while higher accelerations are indicative of normal fetal health

Also, Point plots show that
1. High baseline values of about 142 are indicative of suspect fetal health
2. High accelerations values of about 0.0040 are indicative of Normal fetal health

It was observe that the ground true (label) data is grossly imbalance in favour of Normal(1.0) fetal health. However since the values are ordinal categorical values, focus will not just be on predication accuracy but shall take into account Recall, Precision and F1 Scores of the models built.

3.3 Building and Training the Models

Model building is both an art and a science and as such, deciding on which algorithm to choose from when building a model is rather intuitive and requires experience. Clearly an algorithm meant for clustering cannot be use for supervised learning but beyond that, its more of repeated trials and assessing performance and hyperparameter tuning to see which produces the best result for a given dataset.

3.3.1 Cross Validation

Dataset was split into percentages of training(70), validation(15) and testing(15) and a decision on 5 different classifiers was made to see how they perform on data-set using a stratified 10-fold cross validation.
The choice of 10-fold is in pursuit of a sweet spot which reduces training time without compromising accuracy. Cross validation is one of the best ways to compare various models.[9] An odd number 5 chosen as the number of classifiers was to reduce the probability of a tie when voting. A tie can still occur if one of the models goes completely off the track in voting and the other 4 split their votes equally. In that case, any of the highest vote will be selected.

The choice algorithms are :-
Random Forest
Support Vector Classifier
Decision Tree
Ridge Classifier
K-Nearest Neighbour

After the training, all models performed excellently with around 90 percent on cross validation accuracy. Random forest however had a bit of an edge with 94 percent. This is closely followed by Decision tree (91 percent), Support Vector Machine and KNN at 90 percent and lastly Ridge Classifier at 86 percent
With the models built, the project proceeded to investigate what the result on each model will be, using the 15 percent validation data earlier set aside. Again, all models performed excellently.
Again, Random Forest out performs all the other models on all metrics of F1 score, Precision, Recall and Accuracy

3.3.2 Ensemble Modeling performance

Ensemble learning involves combine the predictions of several base estimators built with a given learning algorithm in order to improve generalizability / robustness over a single estimator[10].

Bagging and boosting are two popular types of Ensemble learning approach. Bagging involves getting central value from the predictions of the models. This could be mean, mode or median[10]. This project employs voting by all five model with the mode selected as the prediction. In a case for example, where a prediction of 1 is made by two models and the other three models predict 2, then the Ensemble model will output 2 as the prediction because 2 is the modal prediction.

Test data-set was feed to all five models and Ensemble prediction was made. The Ensemble prediction performance fell marginally lower than that of Random forest. The performance metric score stood around 93 percent on F1 score, Recall, Precision and Accuracy

4 Results

THE results obtained during this project on validation data are quite interesting. As stated previously, cross validation accuracy scores on all five models yielded the following

MODELS	CV. ACC.
1. Support Vector Classifier (SVM)	0.902558
2. Random Decision Forest (Rand)	0.938867
3. Ridge Classifier (Ridg.)	0.860226
4. Decision Tree (D.Tr.)	0.914656
5. K Nearest Neighbour (KNN)	0.899900

Table 3: Cross validation accuracy on validation data

While Random forest takes the lead, ridge classifier performed the less. Even at that, a cross validation score of 86 percent is still quite good.

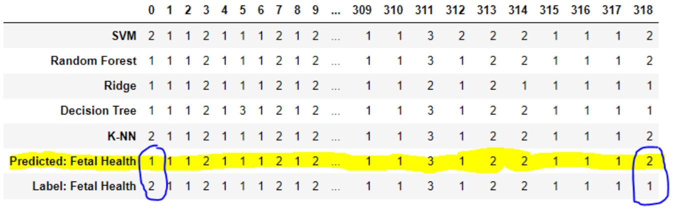
Testing the model on unseen data yields the following

Model	F1 Sc.	Precis.	Recall	Accuracy
SVM	0.905956	0.908146	0.905956	0.905956
Rand	0.946708	0.945821	0.946708	0.946708
Ridg.	0.884013	0.874123	0.884013	0.884013
D. Tr.	0.927900	0.930069	0.927900	0.927900
KNN	0.899687	0.901719	0.899687	0.899687

Table 4: Model performance metrics on test data

All five models were excellent on all performance metric score with regards to test data. However, Random forest wins the trophy, reaching about 95 percent on F1 score

Below is an excerpt from predictions made by the models. The row coloured yellow is the prediction made by the Ensemble model. The output of the Ensemble model is the mode. The first and the last record circled in blue is seen to have been wrongly predicted by the Ensemble model.



	0	1	2	3	4	5	6	7	8	9	...	309	310	311	312	313	314	315	316	317	318
SVM	2	1	1	2	1	1	1	2	1	2	...	1	1	3	2	2	2	1	1	1	2
Random Forest	1	1	1	2	1	1	1	2	1	2	...	1	1	3	1	2	2	1	1	1	2
Ridge	1	1	1	2	1	1	1	2	1	2	...	1	1	2	1	2	1	1	1	1	1
Decision Tree	1	1	1	2	1	3	1	2	1	2	...	1	1	3	1	2	2	1	1	1	1
K-NN	2	1	1	2	1	1	1	2	1	2	...	1	1	3	1	2	2	1	1	1	2
Predicted: Fetal Health	1	1	1	2	1	1	1	2	1	2	...	1	1	3	1	2	2	1	1	1	2
Label: Fetal Health	2	1	1	2	1	1	1	2	1	2	...	1	1	3	1	2	2	1	1	1	1

Fig. 1: Ensemble learning Prediction highlighted in yellow

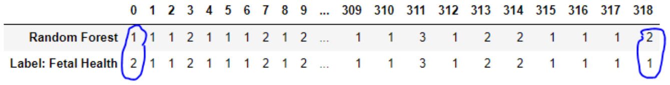
With a score of about 93 percent on all performance metrics, the Ensemble model is wrong 7 in every 100 predictions.

Ensemble Model :

F1 Score	0.9278996865203762
Recall	0.9278996865203761
Precision	0.9266340984689619
Accuracy	0.9278996865203761

Table 5: Performance metrics from Ensemble model

Random forest alone on the other hand gives a slightly higher performance metric score of 94 percent which means Random forest will be wrong in its prediction, 6 out of every 100 times. This is not a significant improvement over the Ensemble model. Moreover, the first and the last record circled in blue is also seen to have been wrongly predicted by Random forest. Random Forest :



	0	1	2	3	4	5	6	7	8	9	...	309	310	311	312	313	314	315	316	317	318
Random Forest	1	1	1	2	1	1	1	2	1	2	...	1	1	3	1	2	2	1	1	1	2
Label: Fetal Health	2	1	1	2	1	1	1	2	1	2	...	1	1	3	1	2	2	1	1	1	1

Fig. 2: Random Forest Model Prediction

F1 Score	0.9404388714733543
Recall	0.9404388714733543
Precision	0.9394874393293946
Accuracy	0.9404388714733543

Table 6: Performance metrics from Random forest model

5 Discussions and Conclusions

HAVEN seen the performance of several models on the same fetal-health dataset and the performance of an Ensemble model which upholds the modal prediction of all five models, The project therefore recommends the Ensemble model over the Random forest model since the Random forest model is not significantly higher than the Ensemble model in performance. Moreover as the idiom goes, two 'good' heads are better than one[11]. Therefore a 5-model committee of good models will outperform on the long run a single very good model such as Random forest

One way of improving the performance of the Ensemble model could be to recompute the prediction of the Ensemble model by discarding the prediction of the least accurate model (in this case ridge classifier) and take the average of the remaining four models to the nearest whole number.

Wale Odunsi
April 14, 2022

Appendix A

References

[1] World Health Organisation (WHO),*Maternal mortality*,https://www.who.int/news-room/fact-sheets/detail/maternal-mortality 19 September 2019.

[2] Rosalie M Grivell, Zarko Alfirevic, Gillian ML Gyte, and Declan Devane,*Antenatal cardiotocography for fetal assessment.*, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6510058/ Cochrane Database Syst Rev. 2015 Sep; 2015(9): CD007863.

[3] Brian A. Magowan MBCHB FRCOG DIPFETMED, *Monitoring of the fetus in labour.*,https://www.sciencedirect.com/topics/medicine-and-dentistry/cardiotocography Clinical Obstetrics and Gynaecology, 2019.

[4] iD Tech, *The 5 different parts of a computer—taking a look under the hood.*, https://www.idtech.com/blog/parts-of-a-computer Jun 11, 2019 3:11 PM.

[5] Adam Augustyn, *software.*,https://www.britannica.com/technology/software Clinical Obstetrics and Gynaecology, 2019.

[6] IBM Cloud Education, *Machine Learning*,https://www.ibm.com/uk-en/cloud/learn/machine-learning 15 July 2020

[7] IBM Cloud Education, *Supervised Learning*,https://www.ibm.com/cloud/learn/supervised-learning 19 August 2020

[8] FILEFORMAT Documentation, *What is a CSV file?*,https://docs.fileformat.com/spreadsheet/csv/ February 2018

[9] Andrew Moore, Carnegie Mellon University, School of Computer Science, *Cross Validation - Tutorial slide by Andrew Moore*,https://www.cs.cmu.edu/∼./awm/tutorials/overfit.html Fri Feb 7 18:00:08 EST 1997

[10] Scikit-learn Documentation, *Ensemble methods*,https://scikit-learn.org/stable/modules/ensemble.html 2011

[11] McGraw-Hill Dictionary of American Idioms and Phrasal Verbs, *two heads are better than one*,https://idioms.thefreedictionary.com/two+heads+are+better+than+one 2002