

Fetal Distress Prediction Based on Cardiotocographic (CTG) Data

ML PROJECT PRESENTATION

Group - 7



INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY **DELHI**



Motivation

- ❖ The number of fetal and maternal deaths every year worldwide is staggering.
- ❖ Cardiotocography -
 - ➔ monitoring technique used to determine a fetus' healthy being
 - ➔ by simultaneously records the fetal heart rate and the mother's uterine contractions.
 - ➔ provides obstetricians with crucial information about fetal state, used to detect abnormal fetal state and movements
- ❖ Drawback - visual inspection of the data is often unreliable.
- ❖ Over 50% of fetal deaths are due to this inconsistency in pattern recognition and failure in receiving a timely intervention.

Motivation

What can be done to combat these challenges introduced due to inconsistent interpretations of CTG as a result of human error?

- ❖ Integrate computerised machine learning methods with obstetrician interpretations for better prediction of fetal state.
- ❖ develop a machine learning model that can identify high-risk fetuses accurately comparable to highly trained medical professionals.

We hope that this would play a significant role in reducing fetal mortality and congenital disabilities globally.

Literature Review

Research Paper 1: [Classification and Feature Selection Approaches for Cardiotocography by Machine Learning Techniques](#)

In this paper[2], both R and Python machine learning techniques are used for performance analysis. Four different types of feature selection based on feature correlations and various models are employed for this study.

Research Paper 2: [Fetal state assessment based on cardiotocography parameters using PCA and AdaBoost](#)

Classification labels used in this paper are normal and pathological. *Principal Component Analysis (PCA)* performs dimensionality reduction and feature selection. It struggles with the problem of outliers as the dataset used has no outlier values, which is not justifiable in the real world. [3]

Literature Review

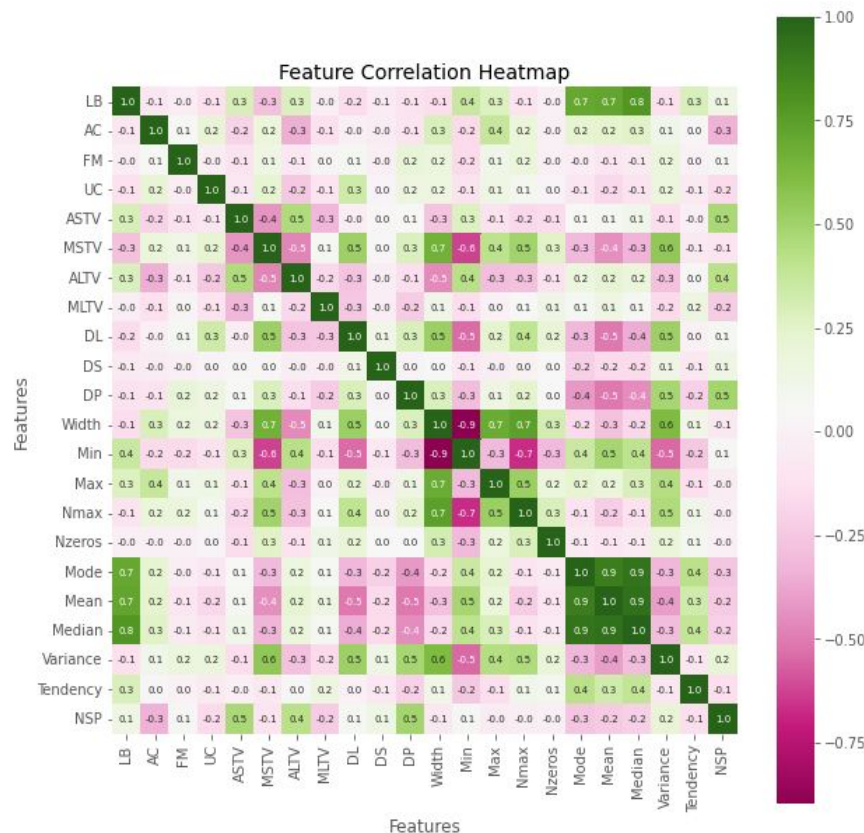
Research Paper 3: [Comparison of Machine Learning Techniques for Fetal Heart Rate Classification](#)

Unlike other studies, the authors in paper [4] also evaluated extreme learning machines[ELM] algorithm with five different activation functions apart from Random Forest Classifier, Support Vector Machines, Artificial Neural Network, and Radial Based function network.

Dataset Description

- We have used the Cardiocotography raw data from the UCI Machine Learning Repository available at <https://archive.ics.uci.edu/ml/datasets/cardiocotography#>
- The data consists of 2126 data samples and 28 features.
- It gives two types of classifications -
 - 1) Morphologic pattern (10 classes) and
 - 2) Fetal state (3 classes - N, S, P)

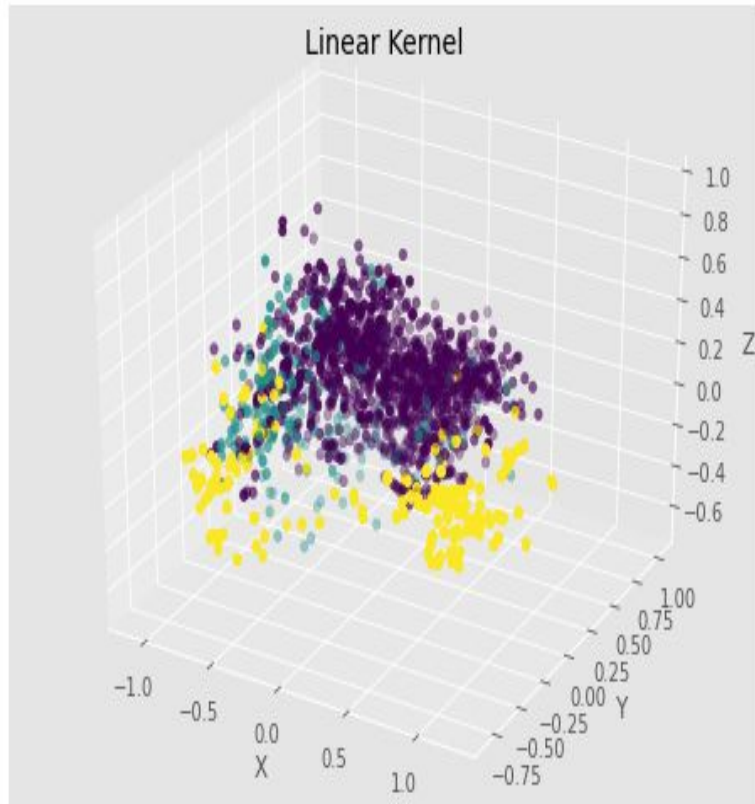
We have plotted a correlation heatmap of various features.



Dataset Description - PREPROCESSING

- 1) **Feature selection and cleaning** - We have dropped 7 irrelevant features as well as rows with null values and removed duplicate data samples. We thus obtained 2112 data samples and 21 features.
- 2) **Normalization** - In order to give equal weights to each feature in the dataset so that no single variable steers model performance in one direction, we performed data normalization using MinMax Scaler to fit between 0 and 1.
- 3) **Oversampling data** -
 - ❖ The division of data into the three classes is imbalanced.
 - ❖ We have used the random oversampling technique to avoid overfitting of the machine learning model on skewed classes by increasing the data samples of the classes with minority instances.

Dataset Description - Principal Component Analysis



- ❖ We performed dimensionality reduction using the Principal Component Analysis techniques on various kernels to reduce the dimension of the data.
- ❖ We noticed that our choice of 21 features was appropriate. Interestingly, we observed that each kernel created two distinct groups of "Pathologic" class : one is close to "Suspect" and one is far from it.
- ❖ However, the data is not linearly separable as observed from the plots of various kernels. We have added the plot for the "Linear" kernel (Figure 3) for reference. All other kernels give similar plots.

Methodology

- ❖ The dataset was split into training and testing set using a 70:30 stratified split. After that we performed a 3 fold cross validation and achieved a train:validation:test split of 47:23:30.
- ❖ The following models were implemented on the dataset using the sklearn library-
 - Logistic Regression
 - Naive Bayes
 - Decision Trees
 - Random Forest
 - K-Nearest Neighbours
- ❖ We also performed hyperparameter tuning using GridSearchCV and chose the best model for training and testing.
- ❖ We used matplotlib and seaborn plotting and visualization → histograms and matrices

Methodology

- ❖ We used the following performance metrics to test our models →
- **Accuracy** - measures the overall efficiency of a classifier.
 - **Precision** - ratio of true positives to the total of the true positives and false positives.
 - **Recall** - ability of a classifier to categorize positively labeled data.
 - **F1 score** - harmonic mean between precision and recall, gives good tradeoff between them

Classifier Algorithm	Accuracy	Precision	Recall	F1 score
Logistic Regression	91.72	91.66	91.72	91.68
Naive Bayes	73.14	88.68	73.14	77.45
KNN	95.26	94.96	95.26	95.01
Decision Tree	95.54	95.44	95.53	95.46
Random Forest	97.30	97.21	97.30	97.19

Table 2 : **Evaluation Metrics**

Methodology

Applied grid search technique on each model to find the optimal hyperparameters for it and got following model parameters :

Classifier rithm	Algo-	Optimal Parameters
Logistic Regression		C=1291.54, max_iter=5000, penalty='l2', solver='saga', multi_class = 'multinomial', random_state = 0
Naive Bayes		estimator = GaussianNB(), 'var_smoothing' = 0.18376
KNN		n_neighbors= 2, metric = 'euclidean', weights= 'distance', algorithm = 'ball_tree'
Decision Tree		max_depth = 12, criterion = "entropy", max_features="auto" splitter = "best", random_state=0
Random Forest		n_estimators = 135, criterion = "en- tropy", max_features = 'auto'

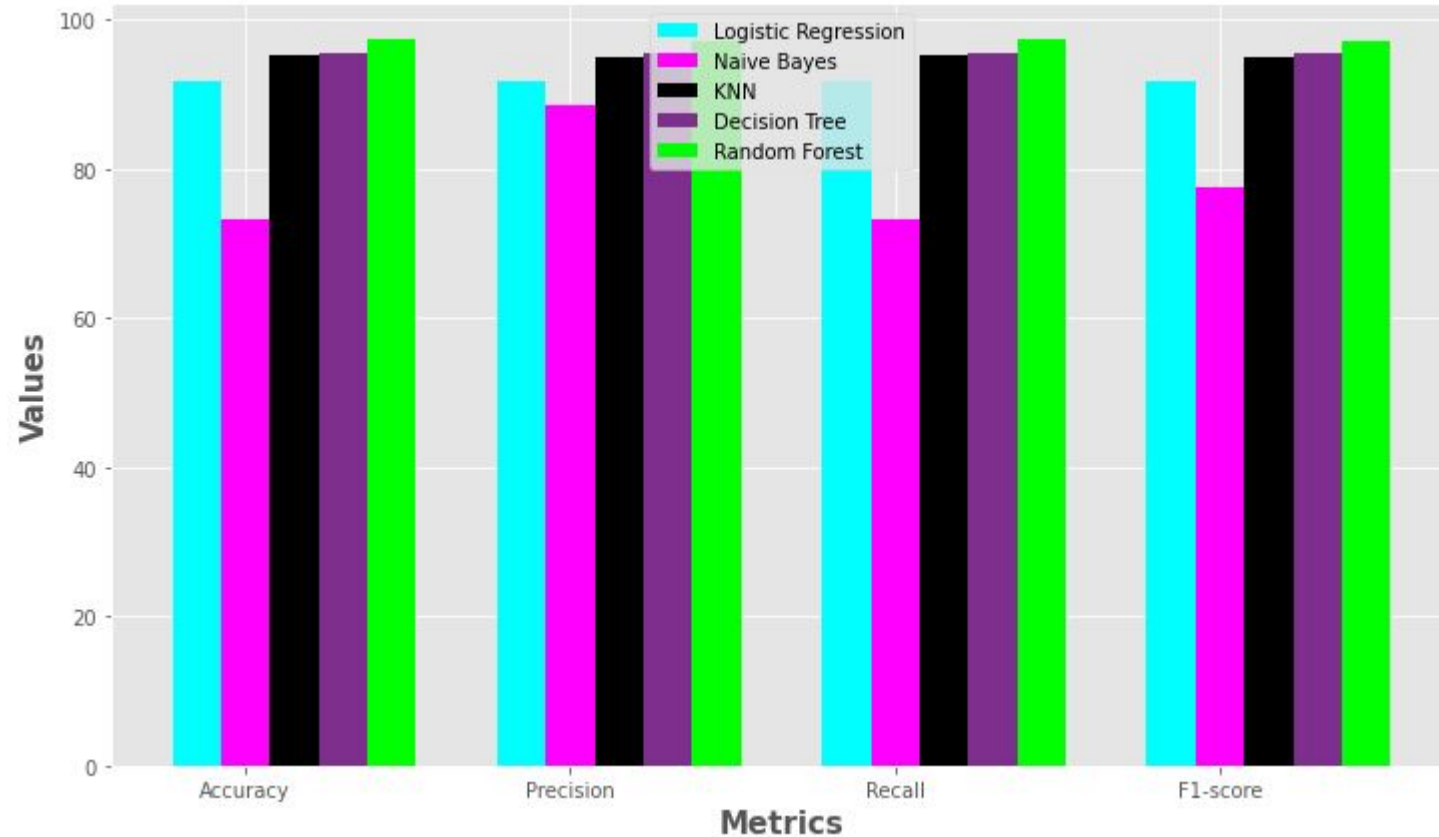
Table 1 : **Tuned Hyperparameters**

Results

The highest accuracy has been observed both in Random Forest (97.30%) and Decision Trees (95.54%) Classifiers followed by KNN (95.26%) and Logistic Regression (91.72%) models, and Gaussian Naive Bayes gives the least accuracy (73.14%).

- ❖ Multinomial logistic regression → perform multi-class classification, which offers better performance over the One-vs-All method.
- ❖ Gaussian Naive Bayes → least accuracy, assumes no dependency between attributes which is not true as the heat map depicts a high correlation between various features.
- ❖ Decision tree → handles high dimensional non parameterized data and works well with non-linearly separable patterns.
- ❖ Random forest → ensemble method, combines the output of multiple unpruned decision trees and makes a prediction based on the majority vote.
- ❖ K-Nearest Neighbours model → works well because it operates on the correlation of features and favours noiseless data

COMPARISON BETWEEN EVALUATION METRICS OF VARIOUS MODELS



Conclusion

- ❖ We learnt that preprocessing, extensive machine learning model techniques, and tools employed play a vital role in classification scores analysis
- ❖ Our methodology not just includes data cleaning and normalization but also incorporates advanced feature selection and engineering methods like Principal Component Analysis used for Dimensionality Reduction.
- ❖ We have achieved almost perfect accuracy, precision, recall, and F1 score metrics of around 95-98% in 3 out of 5 models.

Future Improvement - The dataset does not consider differences in sociodemographic characteristics of pregnant women and some other relevant features like age, nutritional status and so on. The results can be improved by obtaining a larger dataset with diverse sociodemographic characteristics.

Timeline

We are right on schedule in terms of the work done and have covered everything that we needed to do till week 5. Following is the work we have already done -

1	Pre-processing data, Data visualization
2	Feature Analysis and Selection, Plotting Maps, Dimensionality Reduction, Logistic Regression
3	Naive Bayes, Decision Trees
4	Random Forests, K - Nearest Neighbours
5	Analysis of Model Performance, Hyperparameter Tuning

Work left

6	Bagging and Boosting
7	Support Vector Machine, Multi Layer Perceptron
8-9	Analysis of Model Performance, Hyperparameter Tuning Advanced Models, Drawing Final Conclusions
10	Report Writing and Presentation Making

Individual Contribution

- ❖ **Suyashi Singhal** → Data Preprocessing and Visualization, Result Analysis, Training Models and Hyper Parameter Tuning - [DT, KNN, LR], Report Writing, Making Presentation
- ❖ **Ayush Mahant** → Dimensionality Reduction (Principal Component Analysis), Result Analysis, Training Models and Hyper Parameter Tuning - [RF, KNN, LR] Report Writing, Making Presentation
- ❖ **Harshita Gupta** → Plotting Maps, Model selection, Training Models and Hyper Parameter Tuning -[RF, NB, LR] Report Writing, Making Presentation
- ❖ **Rasagya Shokeen** → Plotting Maps, Model Selection, Training Models and Hyper Parameter Tuning - [DT, NB, LR] Report Writing, Making Presentation

These were the assigned responsibilities for each team member. However, all the members equally contributed to all the work done.

THANK YOU

