Comparison of Machine Learning Techniques on Cardiotocography Dataset from UCI Repository

Peter J. Ehmann

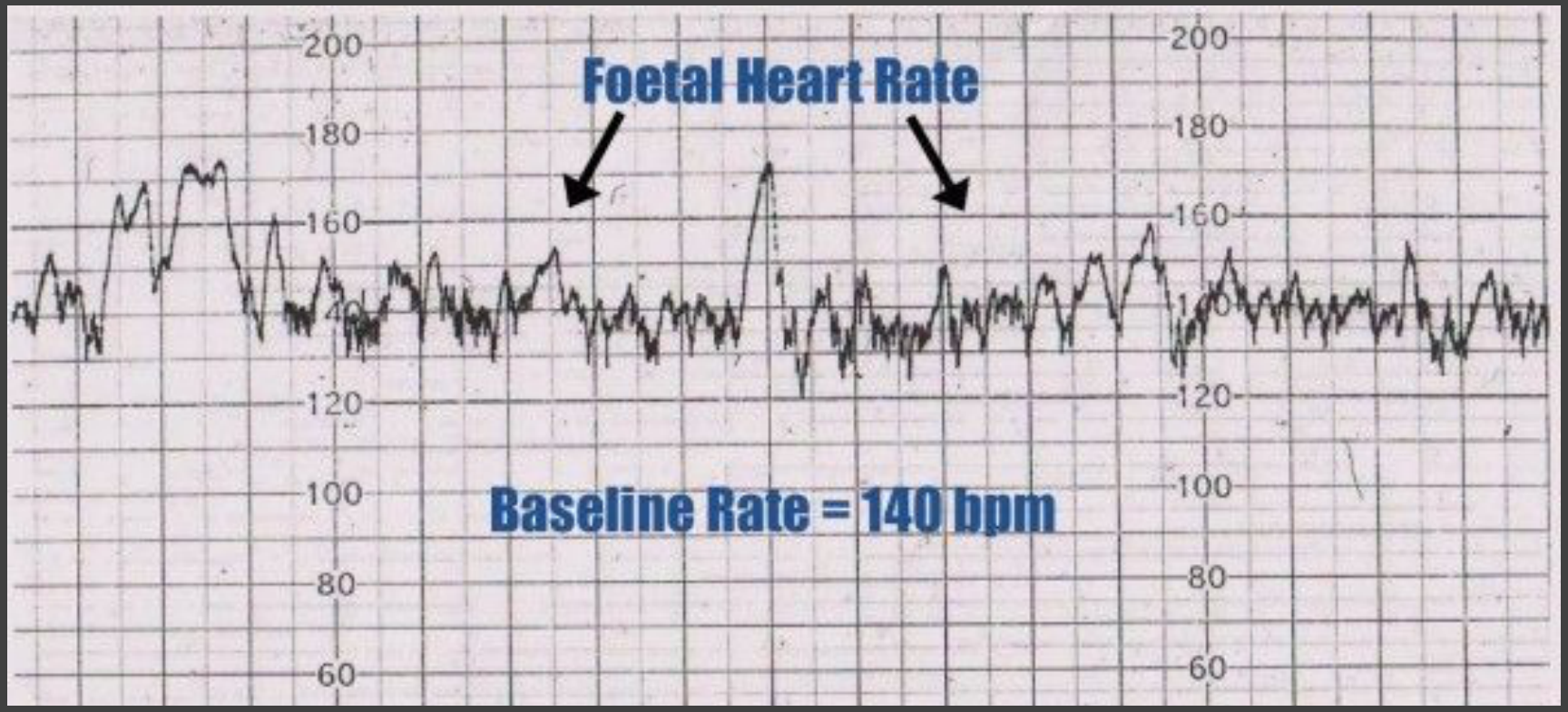
Final Project Presentation

MSDS 534 – Statistical Learning

Rutgers, The State University of New Jersey

Introduction

- Cardiotocography (CTG) – method of assessing fetal heartbeat (ECG) during pregnancy



Dataset

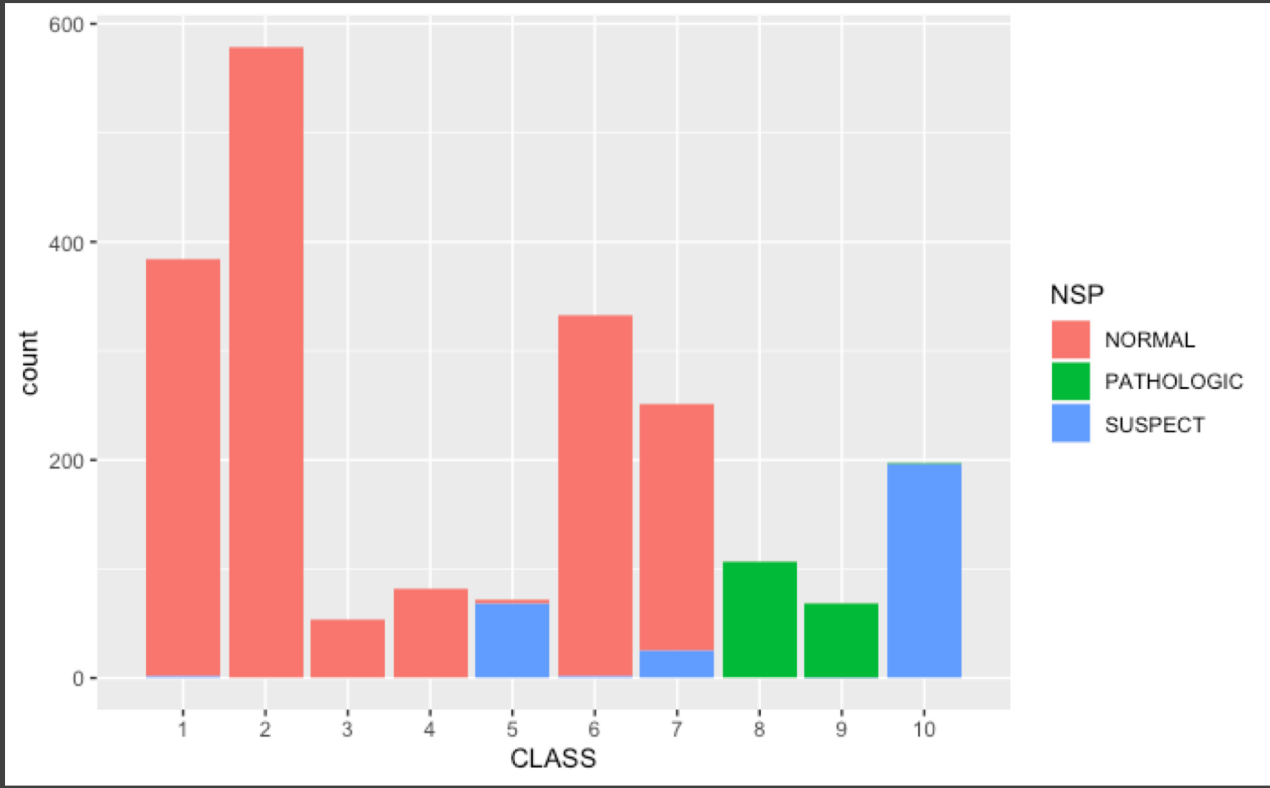
- n = 2126
- p = 21

- Train (60%) = 1275
- Dev (20%) = 425
- Test (20%) = 426

- Minimize false negatives (maximize recall)

1. 10-class (CLASS) vs 3-class Classification (NSP)
2. Coding NSP=2 "Suspect" as *Normal* vs. *Pathological*

Outcome Variables



CLASS	count	percentage
1	384	18.1
2	579	27.2
3	53	2.5
4	81	3.8
5	72	3.4
6	332	15.6
7	252	11.9
8	107	5.0
9	69	3.2
10	197	9.26

NSP	count	percentage
1	1655	77.8
2	295	13.9
3	176	8.3

Methods

- Compare supervised machine learning models and identify best model hyperparameters with grid search

1. Logistic Regression *(glmnet)*
2. Naïve Bayes *(naiveBayes)*

3. Random Forest *(randomForest)*
   - ❖ mtry & nodesize

4. Neural Network *(nnet)*
   - ❖ nnodes & decay

5. Boosting *(gbm)*
   - ❖ shrinkage & n.minobsinnode

Results

Part 1. 10-class (CLASS) vs 3-class Classification (NSP)

	LogReg	NaiveBayes	RandForest	NeuralNet	Boosting
CLASS (10)	0.6371	0.6738	0.8273	0.6521	0.8280
NSP (3)	0.8101	0.7639	0.9035*	0.7200	0.8835

Part 2. Coding NSP=2 "Suspect" as *Normal* vs *Pathological*

	LogReg	NaiveBayes	RandForest	NeuralNet	Boosting
NSP (3)	0.8101	0.7639	0.9035	0.7200	0.8835
S -> N (2)	0.8095	0.7380	0.9523*	0.8095	0.9047
S -> P (2)	0.7623	0.7029	0.8613	0.6831	0.8316

Conclusion

- Models performing best on CTG data (best to worst)

  1. Random Forest
  2. Boosting
  3. Logistic Regression
  4. Neural Network
  5. Naïve Bayes

- Small dataset (*n* = 2126)
- Unequal distribution of classification samples
- Computational limitations & grid search choices