Classification of Fetal Health Data

INTRODUCTION TO MACHINE LEARNING: SUPERVISED LEARNING

Problem Statement

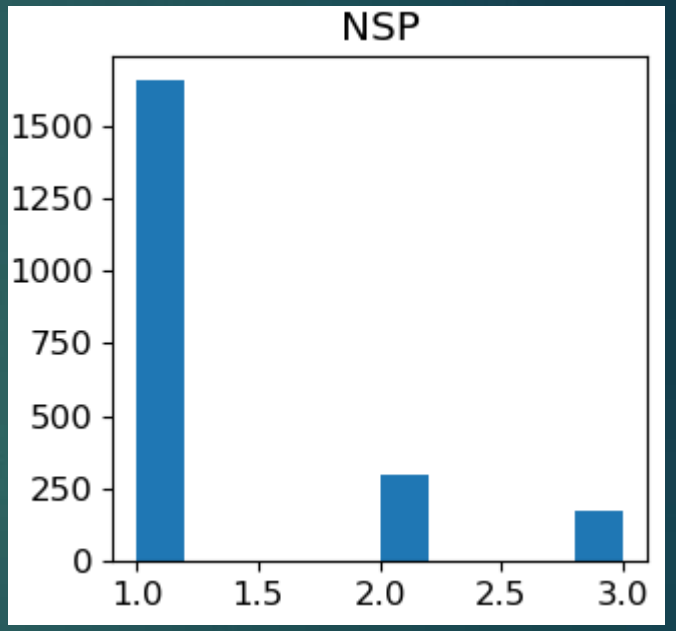
- This project endeavors to apply machine learning algorithms to a dataset of fetal health data to determine whether a given pregnancy should be classified as Normal, Suspect, or Pathologic. If one can easily identify whether certain data indicates an unhealthy fetus, then one can intervene quickly to treat the underlying illnesses and increase the chances of a healthy delivery.

Data

- The data for this study was derived from Cardiotocograms, a simple and inexpensive method of accessing fetal health measures, such as heart rate and fetal movement. This classification problem is important because of the elevated rates of child mortality in many parts of the world.
- This data was first published in 2000 as part of the paper *SisPorto 2.0 A Program for Automated Analysis of Cardiotocograms*, and it is freely available through the UCI Machine Learning Repository.
- Citation: *Diogo Ayres-de-Campos et al. (2000) Cardiotocography Data Set. UCI Machine Learning Repository [https://archive.ics.uci.edu/ml/datasets/cardiotocography]. Irvine, CA: University of California, School of Information and Computer Science.*

Data Attributes

- 2,126 objects with 40 features
- 22 features will be retained, while the others will be dropped. The dropped features are medical measurements that are summarized in the target feature (NSP – Normal, Suspect, Pathologic).
- No data needs to be imputed
- However, the data is imbalanced towards the Normal (1) class. Action will be taken to oversample the minority classes (2 and 3) as part of this project.



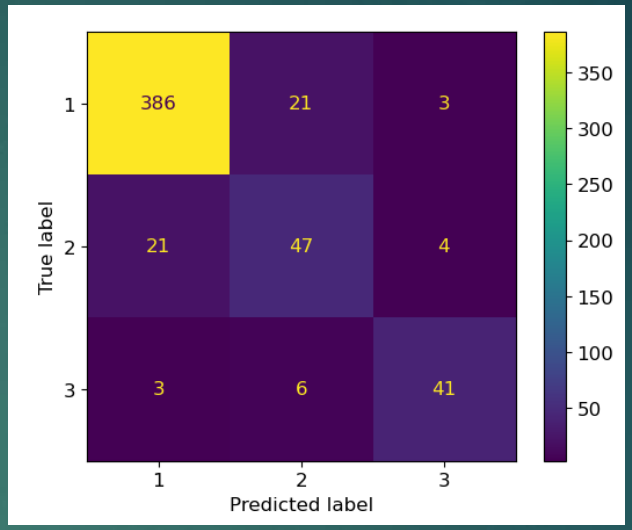
Machine Learning Approach

- K-Nearest Neighbors, Random Forest Algorithm, Support Vector Machines, and Gradient Boosting algorithms
- These will be compared based on error, precision, recall, f-score, and AUC.
- Will use grid search to establish optimal hyper-parameters
- The models will first be trained on the data, then again on the over-sampled data for comparison purposes.
- Models susceptible to collinearity will be trained again on data excluding select correlated features

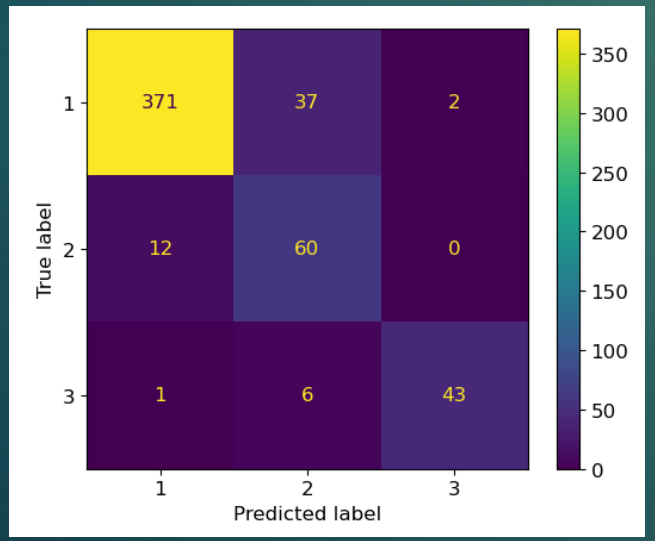
Confusion Matrix

- Gradient Boosting performed the strongest overall

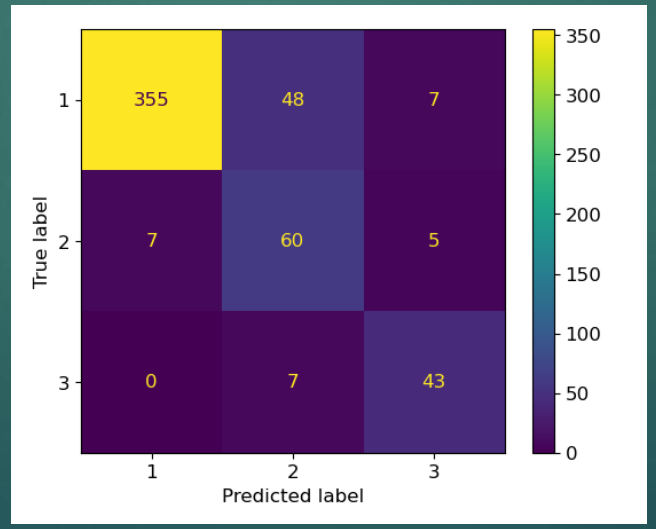
K-Nearest Neighbors



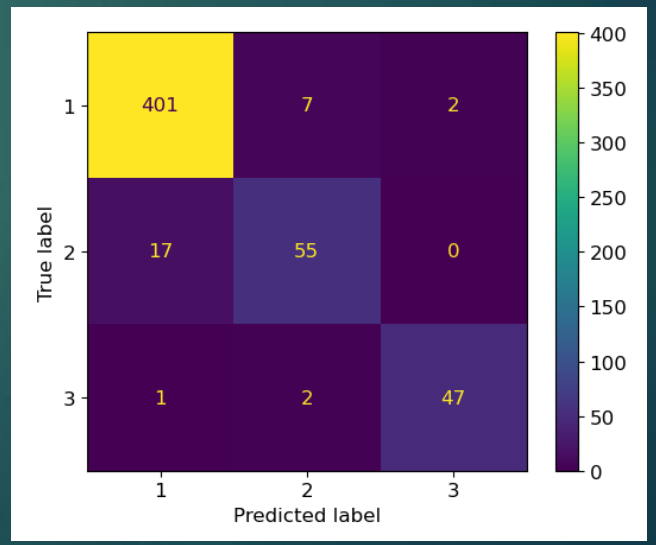
Random Forest



Support Vector Machines

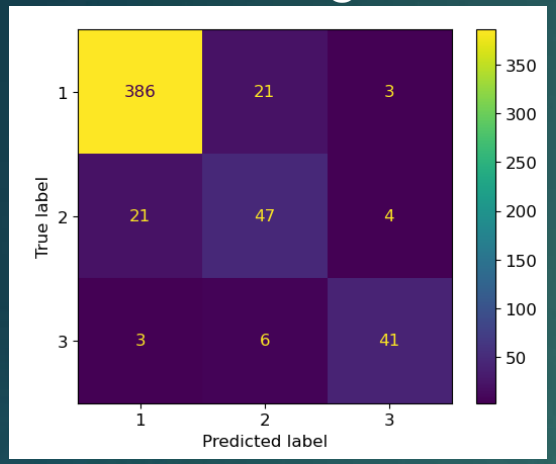


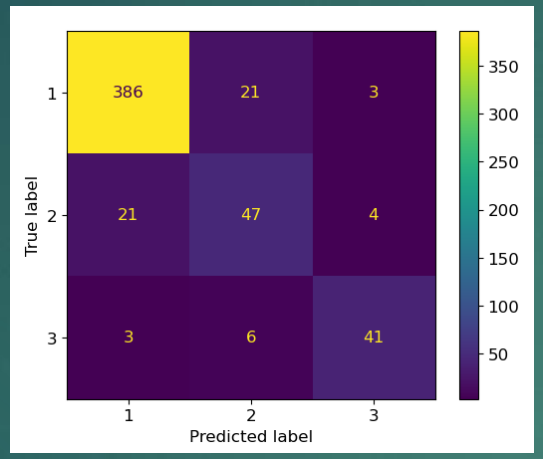
Gradient Boosting



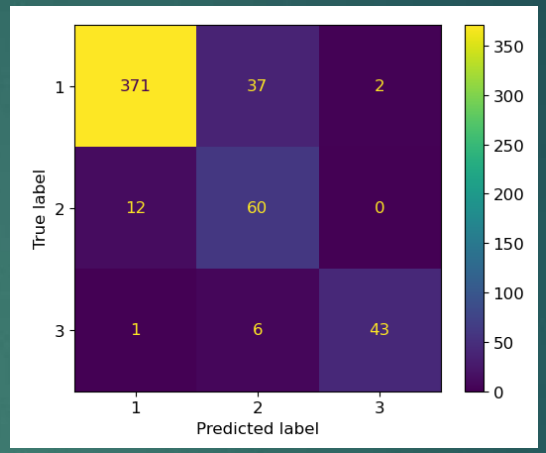
Oversampling

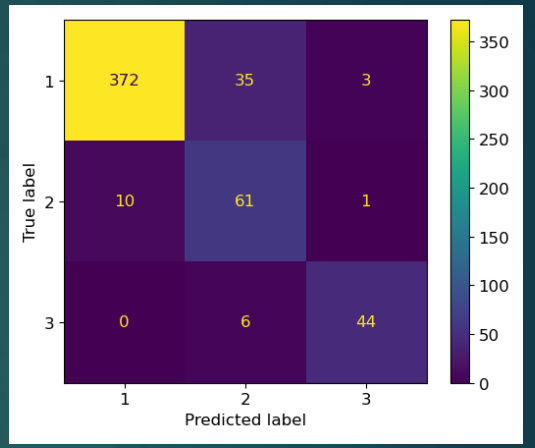
K-Nearest Neighbors



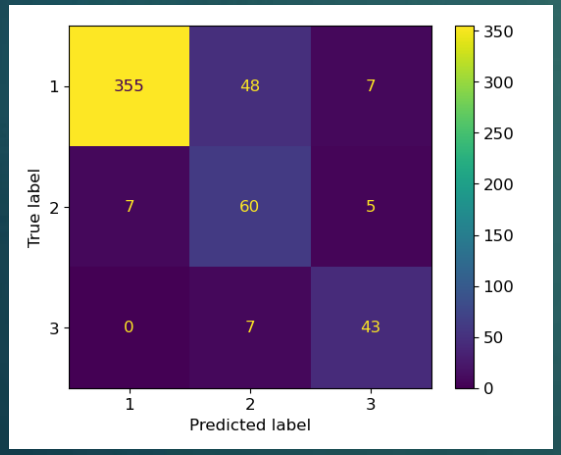


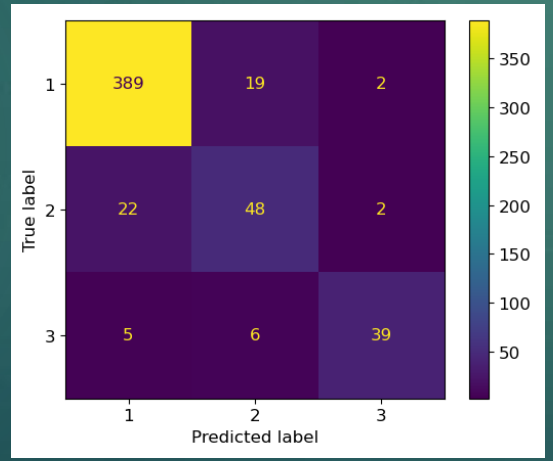
Random Forest



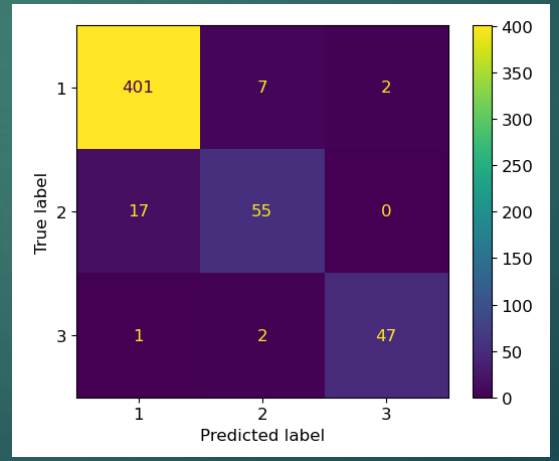


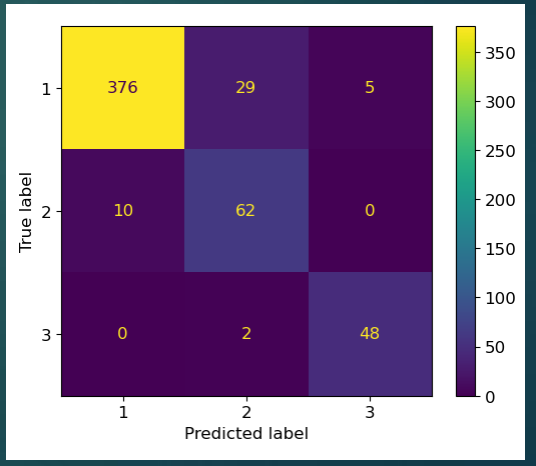
Support Vector Machines





Gradient Boosting

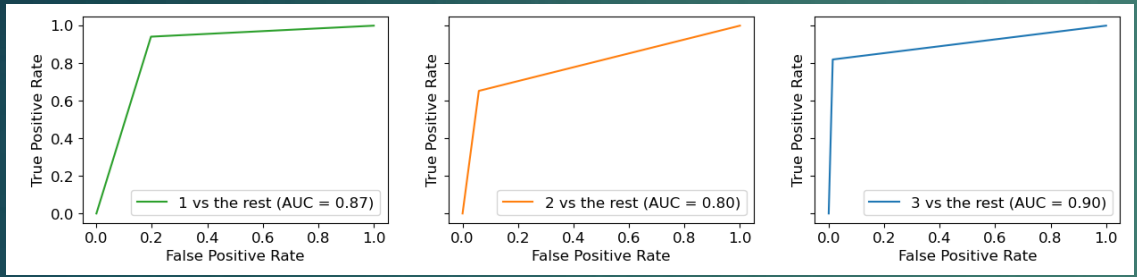




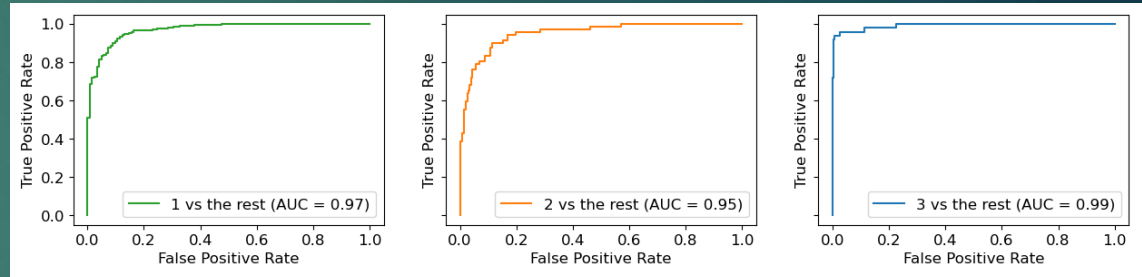
ROC Curve/AUC

- K-Nearest Neighbors struggled with 2 vs Rest, while Gradient Boosting performed strongly across the three classes.

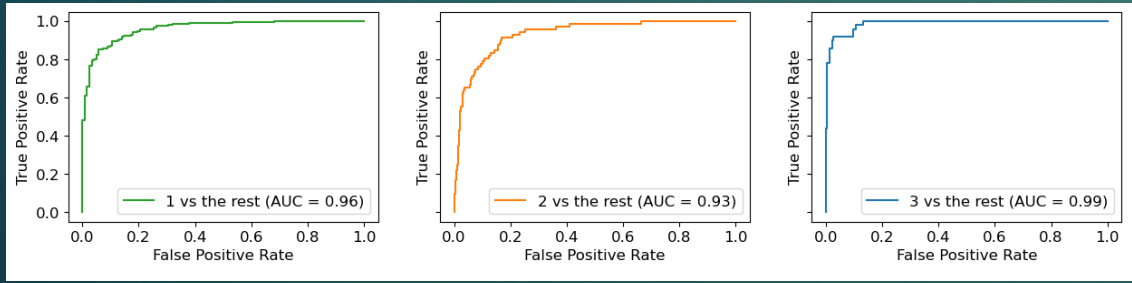
K-Nearest Neighbors



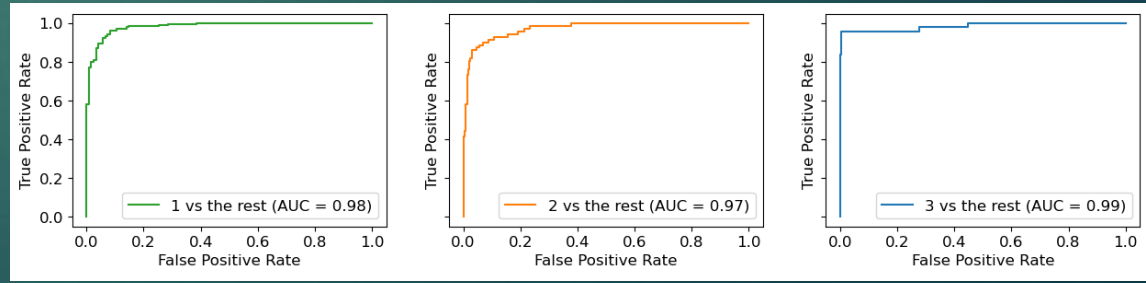
Random Forest



Support Vector Machines

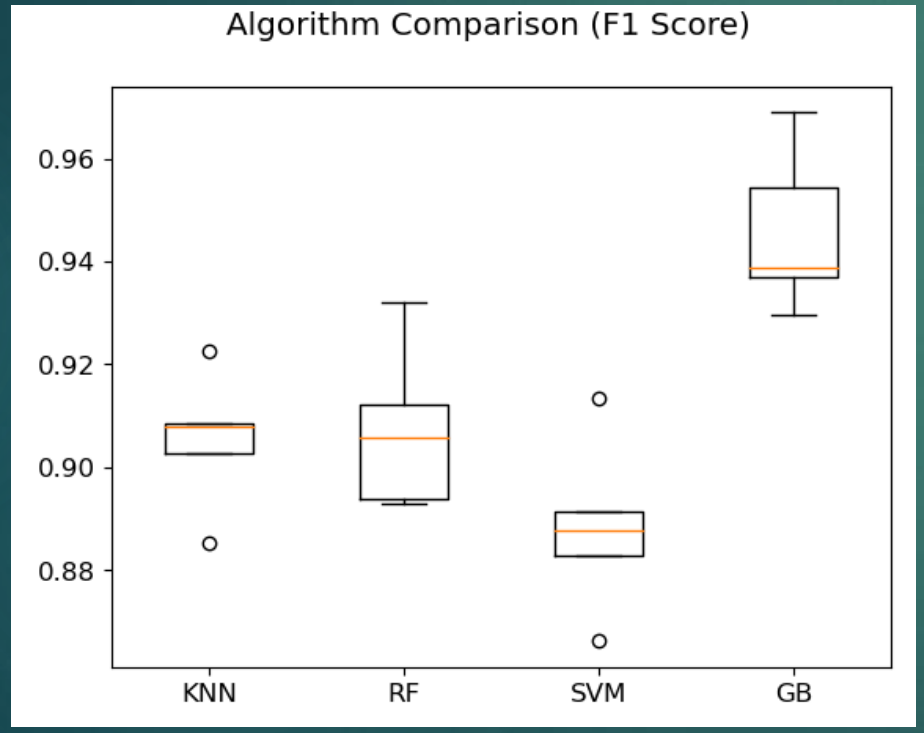


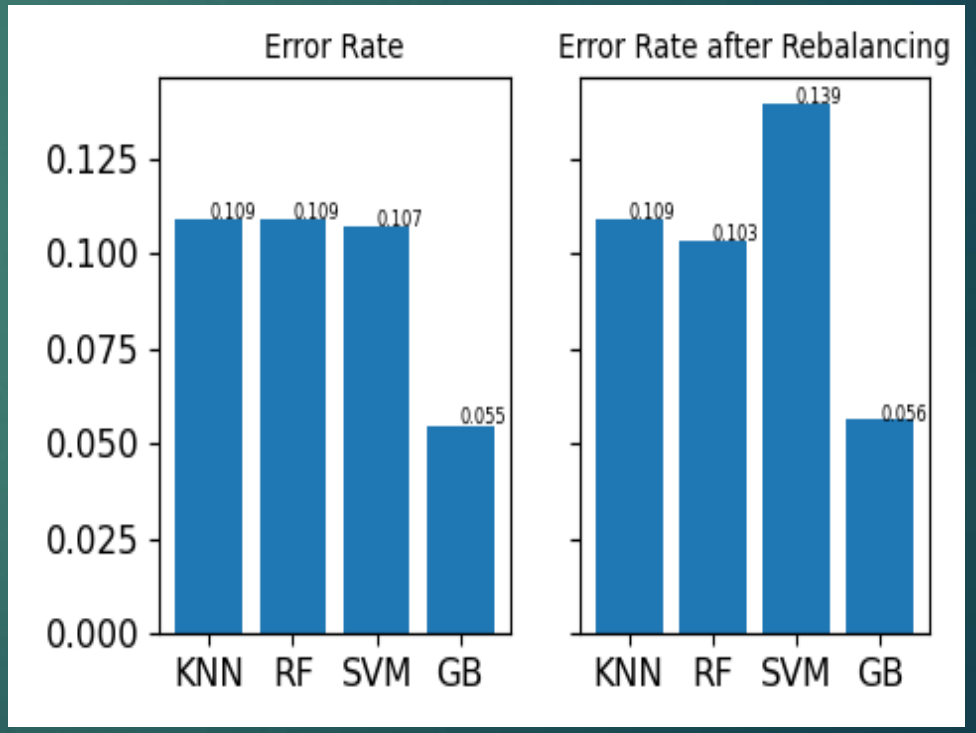
Gradient Boosting



F1 Scores / Error Rates

- Cross validation shows that Gradient Boosting works consistently better throughout the iterations. It also has half the error of the other models.

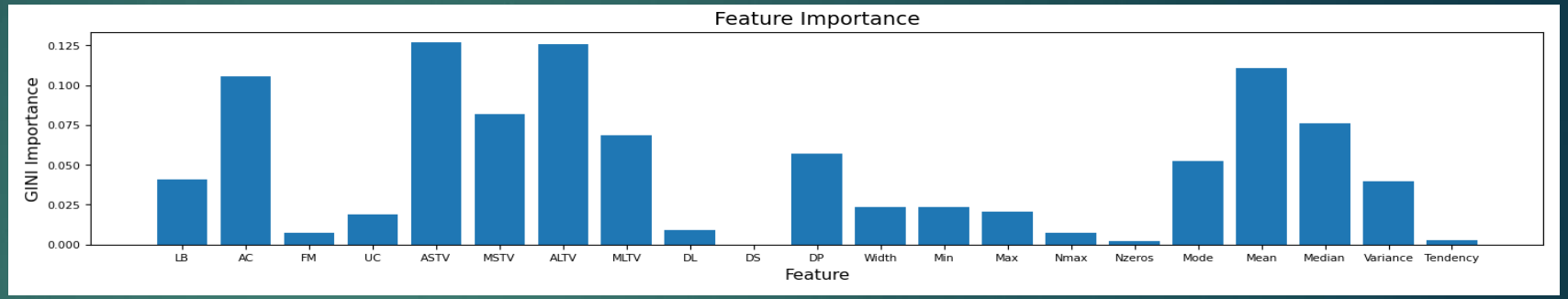




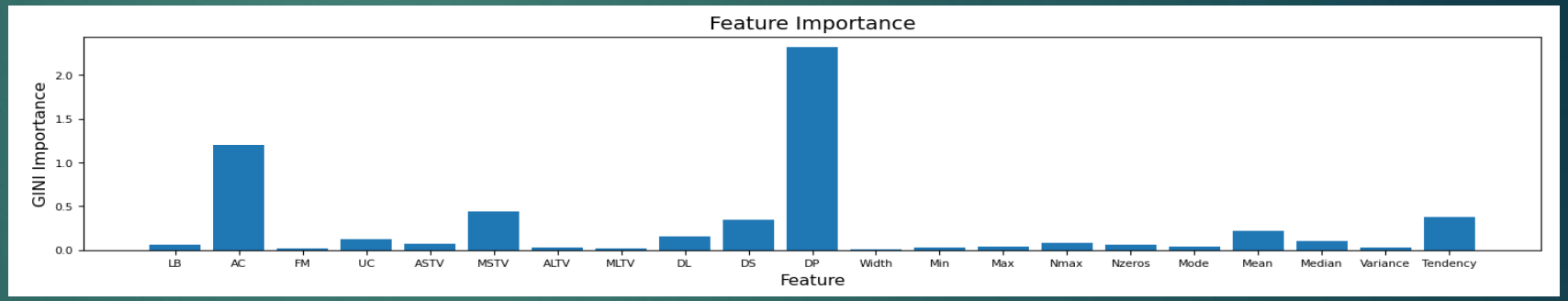
Feature Importance

- ASTV - Percentage of time with abnormal short term variability
- ALTV - Percentage of time with abnormal long term variability
- DP - Number of prolongued decelerations per second

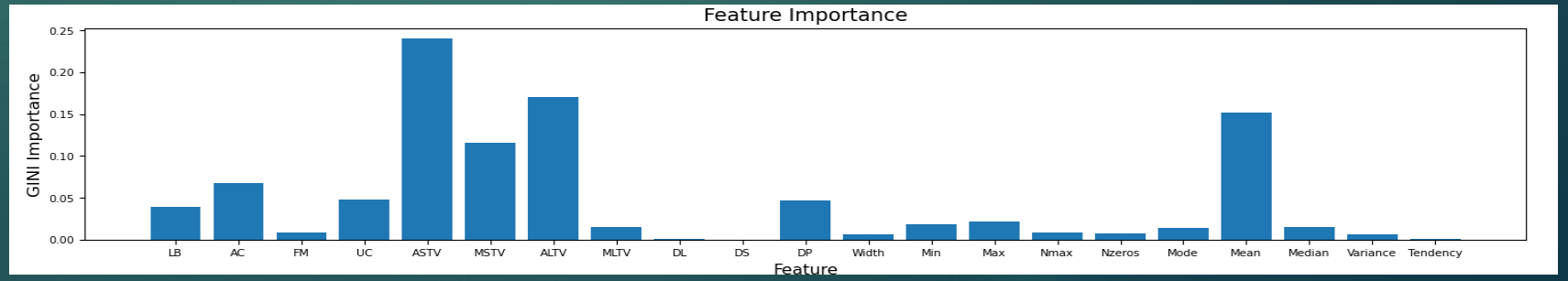
Random Forest



Support Vector Machines



Gradient Boosting



Discussion

- First my assumptions were challenged regarding the need to oversample minority classes of data.
- I also learned to trust the use of grid search. Instead of manually searching for the best hyperparameters for these models, it was easiest to just iterate through a selection of values using cross validation.
- I might attempt to use k-fold cross validation to measure algorithm success in future projects. Segmenting the data into train and test segments is a valid approach. However, using cross validation I would have had more data to train on, and the results would have been averaged over a number of iterations.
- In this project I also learned to trust F-scores and the Area Under the Curve measure. When you have imbalanced data, then you need to have a good measure of overall performance across the classes.

Conclusion

- While all four algorithms performed well in this task, Gradient Boosting achieved the highest marks. Random Forest performed second best, while the K-Nearest Neighbors and Support Vector Machine algorithms performed least successfully.
- Overall, if one can easily identify whether certain Cardiotocogram data indicates an unhealthy fetus, then one can intervene quickly to treat the underlying illnesses and increase the chances of a healthy delivery. From this investigation, it is clear to me that a classifier based on the Gradient Boosting can ensure that good predictions are made about whether medical intervention is needed during a pregnancy.