Predicting Fetal Health Classification

Introduction

We are interested in predicting fetal mortality because it is a current and pressing issue. We want to apply the skills we have learned in class to a real-life situation to truly see the value and importance of machine learning. This dataset gave us the opportunity to explore different models and their prediction ability on a numerical and binary classification dataset. From this dataset, we created our research question: What machine learning model best predicts fetal health status?

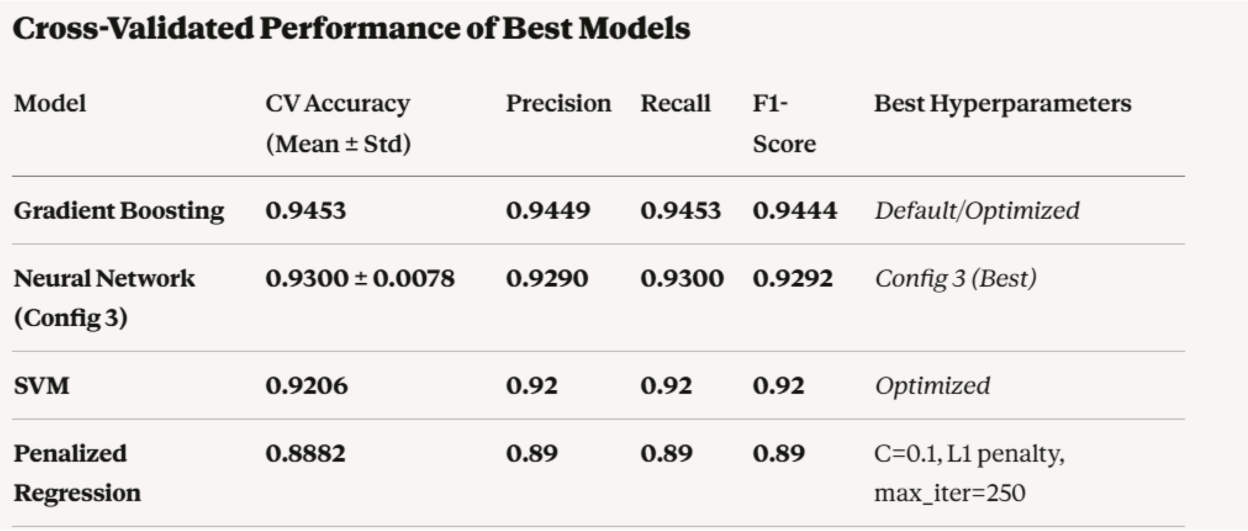
Methods

The dataset we decided to use is the "Cardiotocogram (CTG) dataset", which was originally sourced from the UC Irvine Machine Learning Repository (https://archive.ics.uci.edu/). It has over two thousand (2,126) records from cardiotocogram exams and contains 21 physiological features, but we decided not to use the histogram features (11 total features instead). This dataset also had a target variable (fetal health), which was categorized into "Normal", "Suspect", and "Pathological" classes, which we decided to group into just "Normal" and "Suspect".

The models we used were an SVM, a penalized logistic regression, a gradient boosting, and a neural network model. In all of these models except the neural network, we made a pipeline and implemented a standard scaler to scale the data before making the respective model. Our cross validation strategy was also similar across all models, where we chose several hyperparameters and ran them through grid search to find the best parameters. For the neural net, this was physically accomplished by making a for loop through all of the different parameters. The metrics we used were mainly confusion matrices and f1 accuracy scores due to the fact that these were binary classification models.

Results

We deployed 4 different models- all using cross validation. We deployed a Neural Network, Gradient Boosting, Penalized Regression, and SVM. The accuracy we got from each were 0.93, 0.95, 0.89, and 0.92, respectively. Looking at the Neural Network, it was a good performer! The confusion matrix printed out 70 for False Negatives (incorrectly predicting as class 2 when it was actually class 1. We want to minimize the number of False Negatives because it may be dangerous if patients continue on without being flagged. Gradient Boosting was our top performer and 61 False Negatives and only 32 false positives - demonstrating precision.Penalized Regression had a decent performance with 21 False negatives, but of course a lower accuracy. Finally the SVM had a good performance- with low false positives, but 14 false negatives! This indicates that SVM may be more conservative in predicting class 1. Also, we noted that there was some class imbalance and attempted to combat that with stratified cross validation in some. Just something we kept in mind. Not to mention, From our feature importance analysis, we found that the two strongest predictors of fetal mortality were abnormal short term variability and prolonged decelerations (which could help signal acute distress and moment-to moment changes). Finally, from these results we deployed the Gradient Boosting on the test set.



Cross-Validated Performance of Best Models

Model	CV Accuracy (Mean ± Std)	Precision	Recall	F1-Score	Best Hyperparameters
Gradient Boosting	**0.9453**	**0.9449**	**0.9453**	**0.9444**	*Default/Optimized*
Neural Network (Config 3)	**0.9300 ± 0.0078**	**0.9290**	**0.9300**	**0.9292**	*Config 3 (Best)*
SVM	**0.9206**	**0.92**	**0.92**	**0.92**	*Optimized*
Penalized Regression	**0.8882**	**0.89**	**0.89**	**0.89**	C=0.1, L1 penalty, max_iter=250

Discussion

All of our models performed relatively well when tested on the training set. Every accuracy-f1 score was above 90%, with our Gradient Boosting model performing the best at 95%. The accuracy-f1 score of the Gradient Boosting model increased to 99.5% when tested on the test set. These findings suggest that if this model was used on other fetal health data that it has never seen before, it will be correct almost 100% of the time. We are confident that this model works very well, but is not overfitting because the accuracy-f1 score on the training set was not higher than that of the test set.

While each model performed well, Gradient Boosting performed better than the others because of the non-linearity of the features. The decision trees that it builds and uses to predict are very accurate on non-linear features, and are able to capture these patterns more effectively than the other models, which are better suited for linear patterns. Even though our final model was extremely accurate, there are still some limitations to this project. The main limitation is that our model may not be applicable to all areas of the world because of differences in culture and healthcare availability that could strongly affect the features the model uses. Certain patterns that our model is finding in this dataset might not be evident in other fetal health datasets.