Predicting Fetal Heart Health with Supervised Machine Learning

Holly Figueroa and Karlie Schwartzwald

College of Science and Technology, Bellevue University

DSC 630: Predictive Analytics

Professor Andrew Hua

March 4th, 2023

A Machine Learning Approach to Cardiotocography Interpretation

Cardiotocography (CTG) is a procedure widely used over the course of pregnancy and during labor (*Fetal Heart Monitoring*, 2019). CTG allows monitoring of fetal heart rate and uterine contractions to prevent fetal hypoxia, or oxygen loss in body tissues (Boudet et al., 2020). Effective modeling for CTG exam data for diagnostic tools can provide additional support for medical professionals and the institutions they work for. Here we will review our work to interpret CTG exam using two separate supervised machine learning models. A complement naïve Bayes classifier and a multilayer perceptron classifier were used to predict fetal heart health outcomes, and results were compared. While both offered compelling accuracy, further evaluation demonstrated shortfalls in predicting non-normal CTG data.

The dataset used to train our models was sourced from Kaggle.com. It contained 2126 rows of patient CTG exams with 21 total features. Most importantly, it contained labeled outcomes that could be used as our target for model training. All features were quantitative and continuous while our target variable was nominal. Fetal CTG outcomes were determined by consensus of a 3-person expert panel with labels as 1 = 'Normal', 2 = 'Suspect', and 3 = 'Pathological".

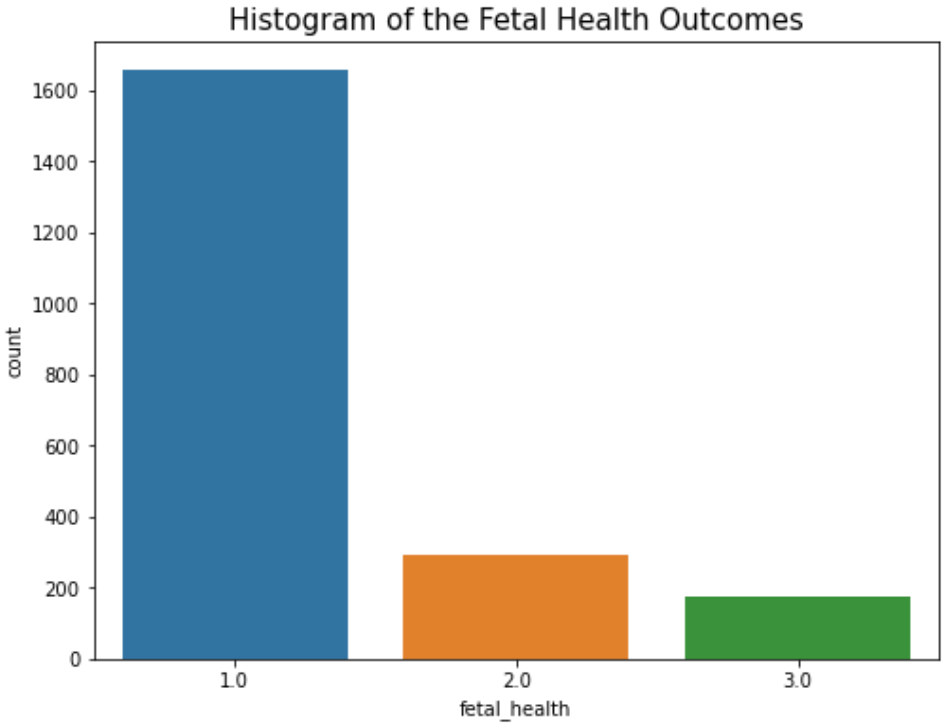
Preliminary Analysis and Data Preparation

The data we used to train our model was without errors and contained a complete dataset free of missing values and duplicates. Histograms and descriptive statistics were used to conduct univariate analysis on all variables. Multivariate analysis was conducted using correlation matrices and box plots to illustrate variable distributions.

Figure 1 shows the distribution of our target feature, and we can see it has imbalanced classes which can cause difficulties for models in making predictions regarding the classes with the fewest instances. You can see in the histogram that for our target feature, the 'Normal' class is considerably larger than the 'Suspect' or 'Pathological' classes.

Figure 1.

Histogram of Fetal Health Outcomes

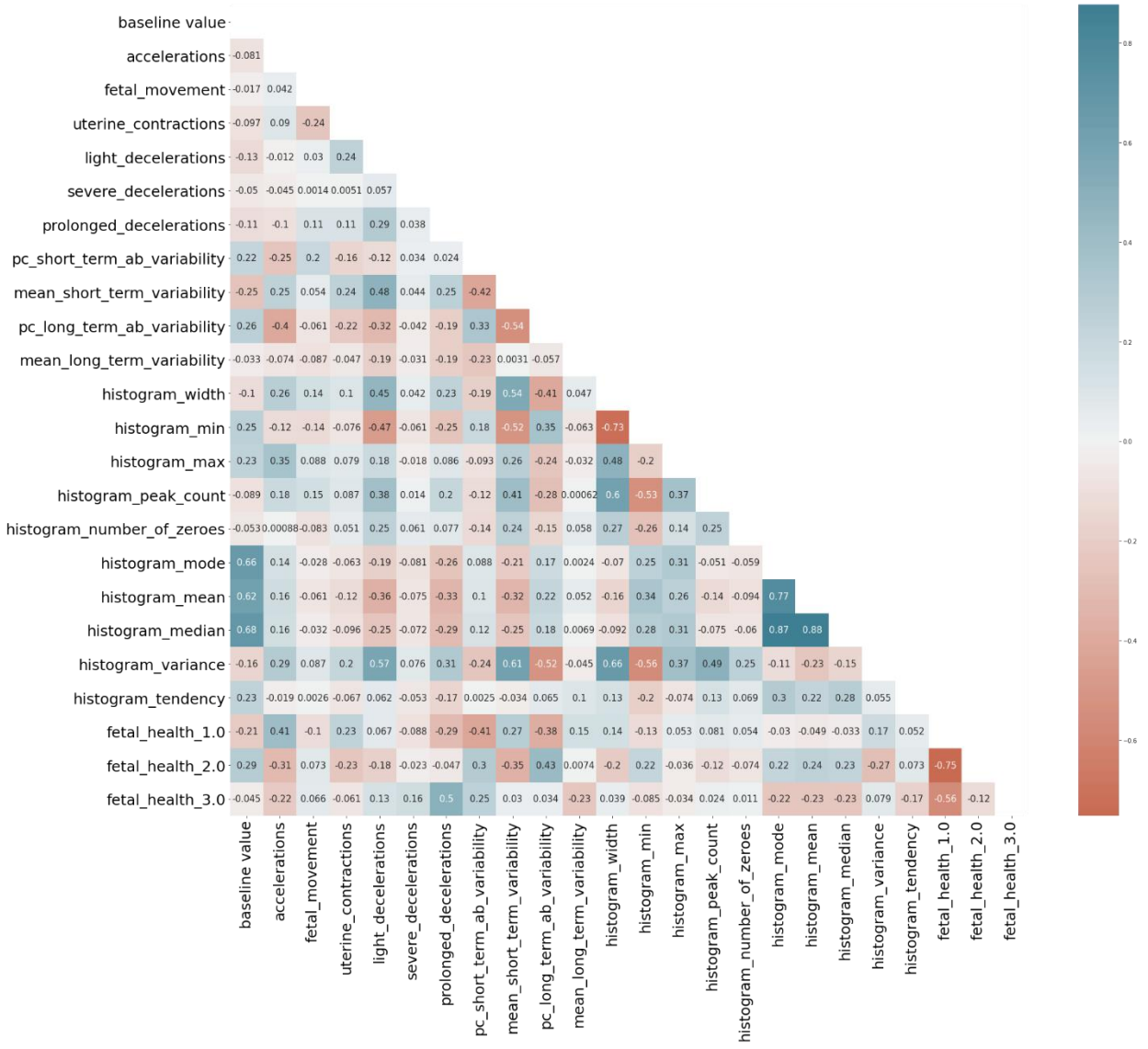


Our preliminary analysis with correlations showed that variation in many of our features are associated with variations in our target classes. The relationships were not necessarily strong, but combined, they appeared suitable for training for our predictive models. In Figure 2, we have a correlation matrix using Kendall's Tau as the correlation measure. The strongest relationships

to our target classes were the features related to CTG involving deceleration at 0.5 Tau and features of abnormal variability, with those Tau's ranging from 0.25 to 0.43.

Figure 2.

Correlation Matrix of Data Features Using Kendall's Tau



Histograms related to heart decelerations per second were found to be highly imbalanced. But upon further inspection using value counts, the rare values represented cases with the most serious health outcome, 'pathological.' Strong correlations between variables were also shown

between different central measures related to ETC histograms (mean, median, and mode), which led to the dropping of the features median and mode.

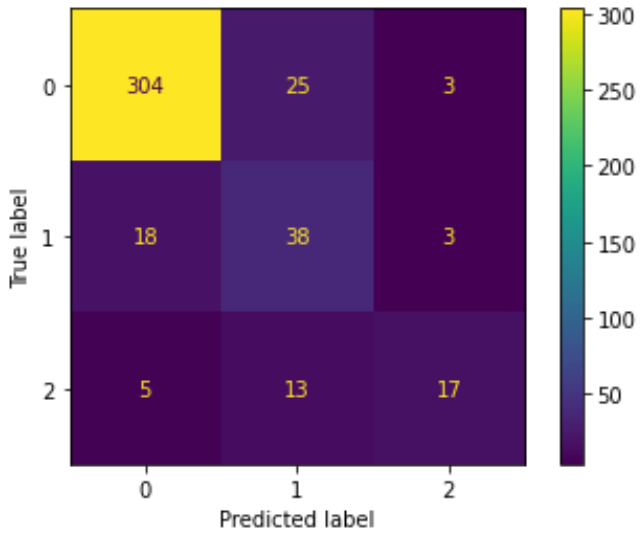
Complement Naïve Bayes Modeling & Evaluation

We chose to use the Complement Naïve Bayes due to its ability to handle imbalanced classes in the target feature, and therefore did not perform manual oversampling on the target feature. We used a train-test-split ratio of 80/20 because it yielded the highest accuracy of any ratio we tested, without overfitting the model too much. We then applied a minmax scalar because the model only accepts positive values as input. After that we did a grid search to perform hyperparameter tuning on the only hyperparameter in a complement naïve Bayes model, using accuracy as our scoring metric.

In evaluation of the complement naïve Bayes model, we primarily used accuracy but also looked more closely at false positive rates, false negative rates, true positive rates, and true negative rates as seen in Figure 3. When we look at the true positive rates for each of the categorical outcomes, we do have some serious concerns. This model was very accurate at predicting healthy outcomes, with a true positive rate of 91%, but much worse at correctly predicting other fetal health outcomes, with true positive rates of less than 65% for each of the other outcomes. The k-fold cross validation of this model had a train score of 0.84 and a test score of 0.84. This could imply some overfitting was happening within this model.

Figure 3.

Confusion Matrix of MLP Predictions vs True Labels for Fetal Heart Health Outcomes



Note: This figure illustrates compares our MLP model's predicted outcome labels for CTG exams against true labels. Here 0 = Normal, 1 = Suspect, and 2 = Pathological.

MLP Modeling & Evaluation

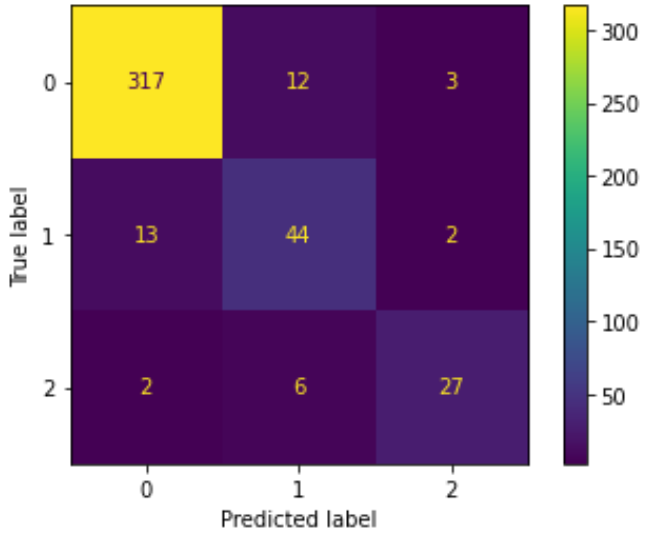
Given that neural networks tend to perform well with larger sample sizes, over-sampling was used to raise sample sizes for our lowest target classes. While this did increase our risk of over sampling, the test split further reduced our small samples for non-normal outcome cases. When completed, we had matching sample sizes across classes for training at 1323 each. Two scalers were tested, and a standard scaler proved to provide slightly higher accuracy. A grid search was conducted along with some manual tuning to find the best performance for our test dataset. The resulting model used a "tahl" activation function and the network was created to have 2 hidden layers of 12 and 6 nodes.

Our MLP model calculated outcomes with a 91% accuracy score. Because this metric reflects overall accuracy and not accuracy across target classes, a confusion matrix was used to compare predictions to true labels. This was used to calculate rates for true positive, true negatives, false negatives, and false positives. Outcomes for 'normal', 'suspect', and

‘pathological’ were returned with true positive rates of 95%, 74%, and 80%, respectively. Of particular concern, given the context of this project, were False Negative Rates for non-normal outcomes. Results returned a FNR of 25% for ‘suspect’ and 20% for ‘pathological’ predictions. While this lent confidence for predicting ‘normal’ outcomes, overall, the results showed the model was not dependable for predicting riskier outcomes.

Figure 4.

Confusion Matrix of MLP Predictions vs True Labels for Fetal Heart Health Outcomes



Note: This figure illustrates compares our MLP model’s predicted outcome labels for CTG exams against true labels. Here 0 = Normal, 1 = Suspect, and 2 = Pathological.

Risks and Ethical Implications

While machine learning healthcare applications (ML-HCAs) and AI recommendation systems can offer powerful insights for the medical field, limits must be acknowledged. The most fundamental limit being correlation is not causation, therefore their role must be additive to care, not a replacement for expertise. A common hope is that the addition of ML-HCAs and AI diagnostics will circumvent bias within the healthcare system, however, these tools are just as

susceptible to introduced bias (Naik, 2022). The data used to train a diagnostic model is critical to how accurately these systems perform for diverse populations. It is unknown how these concerns apply to CTG exams of the unborn. For example, we do not know if CTG exams, in practice, predict or protect all groups equally. We also do not know if the anomalies associated with fetal heart issues present the same across groups. Examining those possibilities carefully is recommended prior to production.

Conclusion

It is worth noting that both of our models performed much better than the null model with regards to accuracy and in that way, they can be considered successful models. When comparing both model results the MLP neural network model performed better with an overall accuracy of 91% compared to our Complement Naïve Bayes accuracy at 84%. Both models, however, under-performed when correctly classifying ‘suspect’ and ‘pathological’ outcomes with false negative rates ranging from 20%-51% across both models. Since the purpose of these models is to identify unhealthy fetal heart health outcomes, we feel that these models are not accurate enough to be deployed for diagnostic purposes. It is possible the imbalance between classes and the nuanced differences between class cases were ill-suited for training these model types. Ideally, a MLP classifier would be trained with a larger sample of non-normal CTG data, as oversampling did not appear to yield comparable accuracy. Without additional samples for training, gradient boosting might also yield better results across our target classes by allowing incorrectly labeled data, to be emphasized.

References

Boudet, S., Houzé l'Aulnoit, A., Demailly, R., Delgranche, A., Peyrodie, L., Beuscart, R., & Houzé de l'Aulnoit, D. (2020). A fetal heart rate morphological analysis toolbox for MATLAB. *SoftwareX*, *11*, 100428. https://doi.org/10.1016/j.softx.2020.100428

Fetal Heart Monitoring. (2019, August 14). Johns Hopkins Medicine. https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/fetal-heart-monitoring

Naik, N. (2022, March 14). *Legal and Ethical Consideration in Artificial Intelligence in Healthcare: Who Takes Responsibility?* Frontiers. https://www.frontiersin.org/articles/10.3389/fsurg.2022.862322/full

Simfukwe, M., Kunda, D., & Chembe, C. (2015). *Comparing Naive Bayes Method and Artificial Neural Network for Semen Quality Categorization.* International Journal of Innovative Science, Engineering & Technology, 2(7) https://ijiset.com/vol2/v2s7/IJISET_V2_I6_90.pdf