FETAL RISK CLASSIFICATION USING CTG DATA: A MACHINE LEARNING APPROACH

Edith Gómez
Sarah Peña
Elier Fajardo

March 2024

PROJECT OVERVIEW

Reducing child mortality rates stands as a paramount objective within the global health agenda. By 2030, nations aspire to eradicate preventable deaths among newborns and children under the age of five. Concurrently, maternal mortality remains a concerning issue, claiming 295,000 lives during and post-pregnancy in 2017.

The convergence of health complications during gestation as a global concern underscores the urgency for innovative solutions. Machine learning (ML) algorithms offer promising avenues for predicting fetal health based on cardiotocographic (CTG) data, categorizing health states into normal, needing assurance, or indicative of pathology.

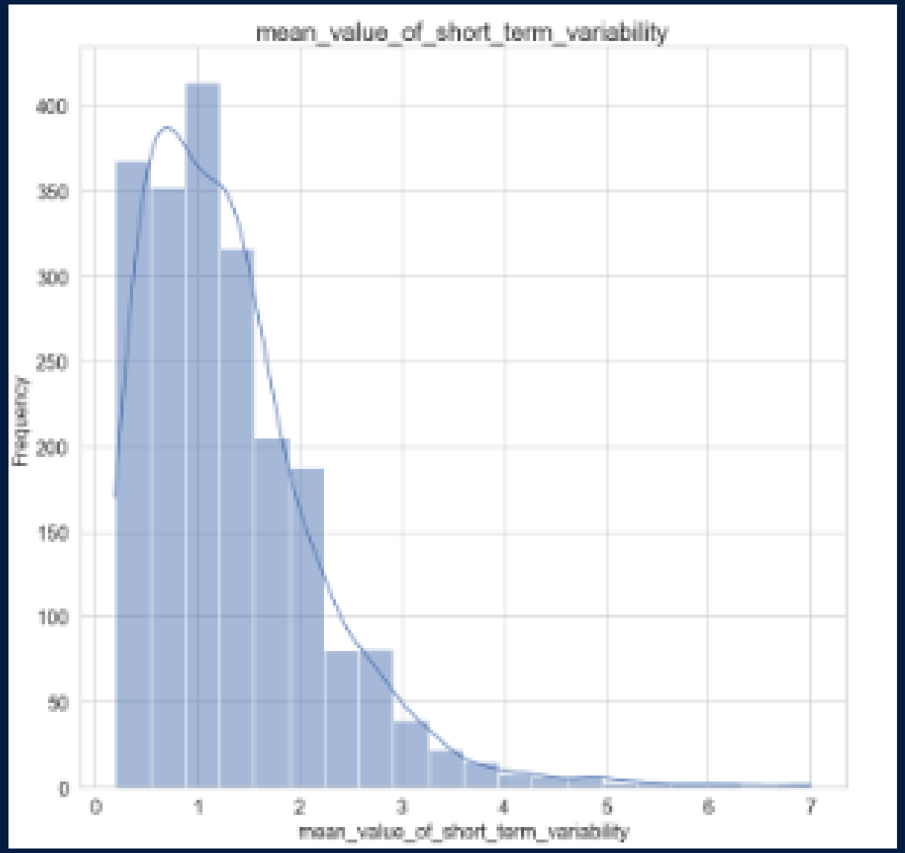
PROJECT PIPELINE

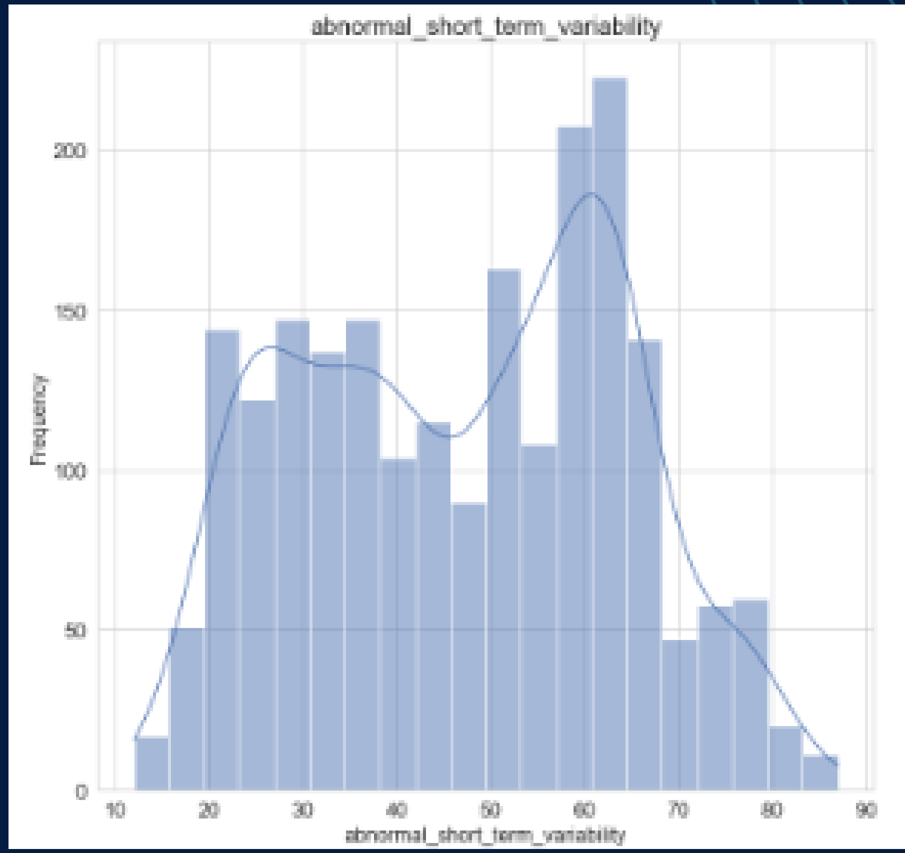
- **Data Collection**: Kaggle dataset repository
- **Data Preprocessing**: Handling of data duplicates, scaling
- **EDA**: Visualization of data distribution, variable correlation, class balance, class differences respect to other variables.
- **Model Training and Evaluation**: Four models were trained and evaluated: dummy classifier, logistic regression, random forest, and a neural network.
- **Conclusions**: Model performance comparison and insights obtained from the experiments.

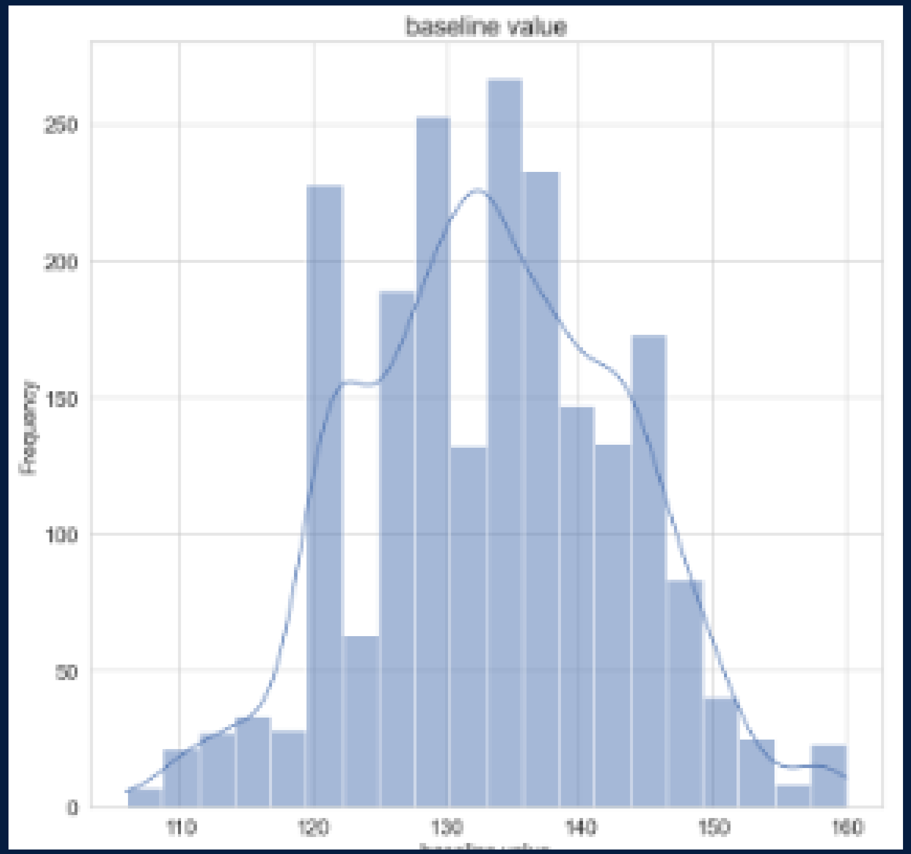
Exploratory Data Analysis

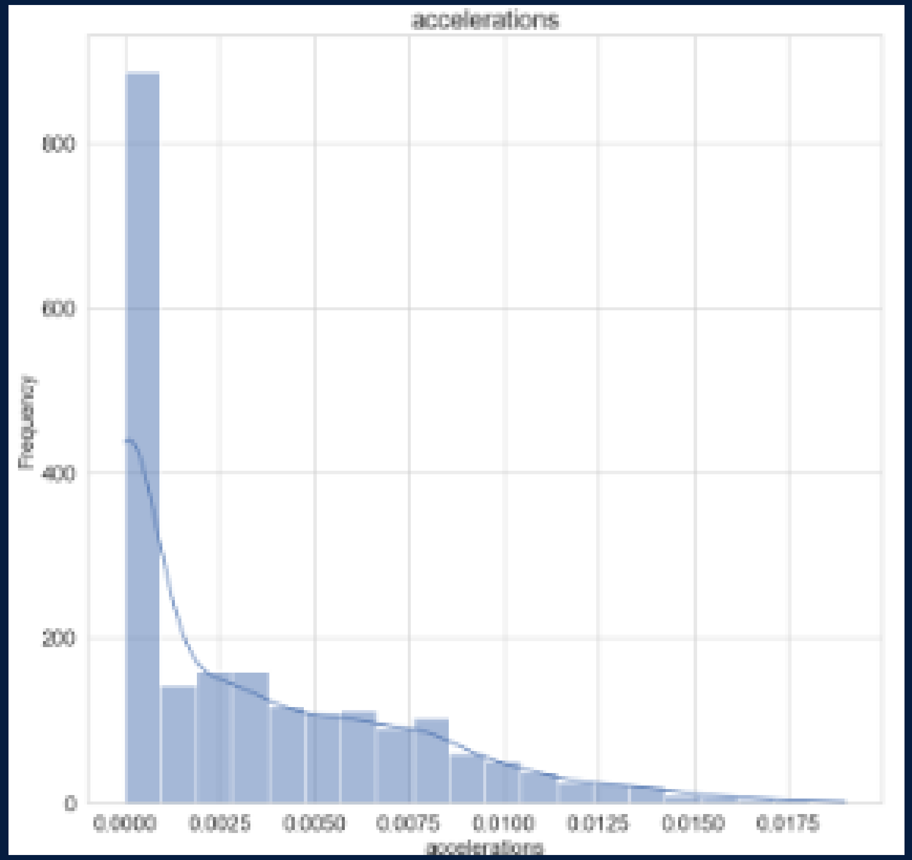
- In first place, we observed the distribution of the variables of the dataset. We idenfitied that some features present skewness in their distribution, others show a normal-like distribution.

- To deal with these differences, we performed data scaling.



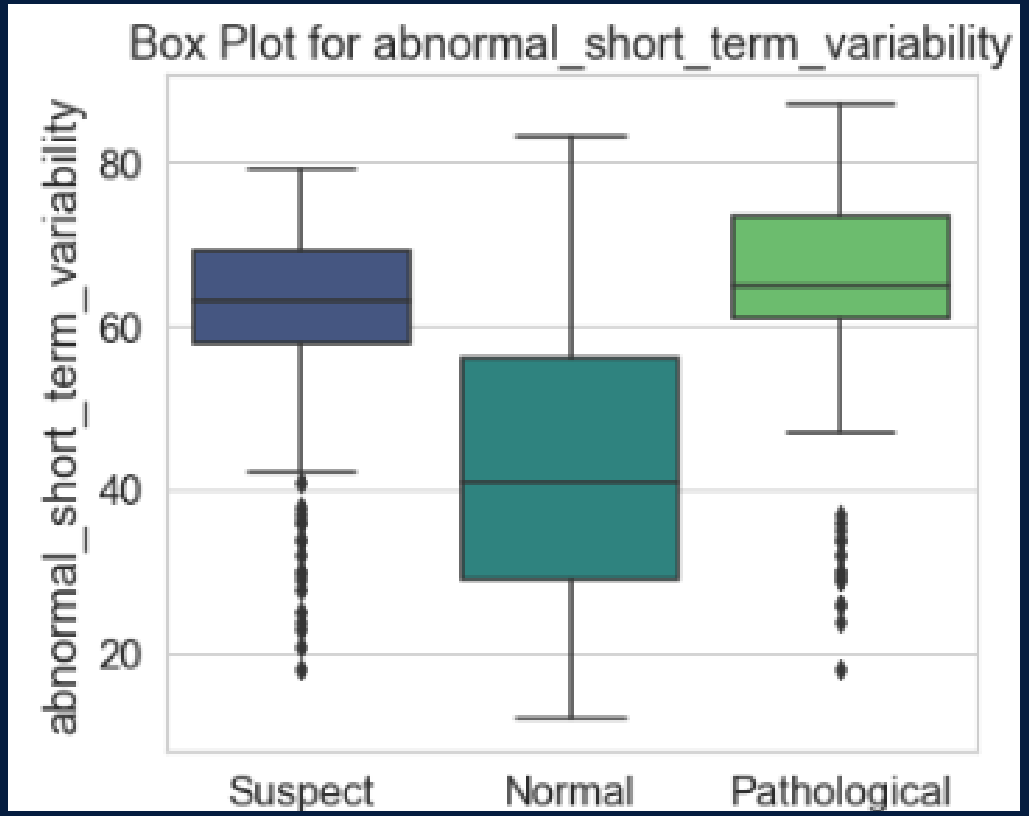


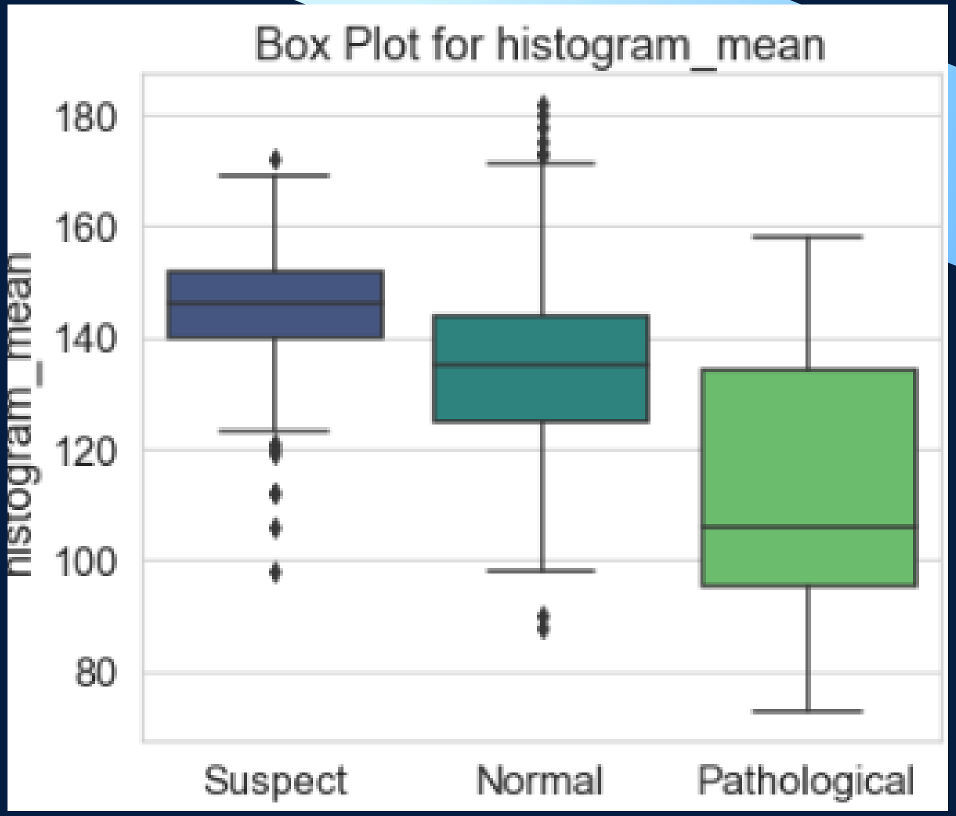


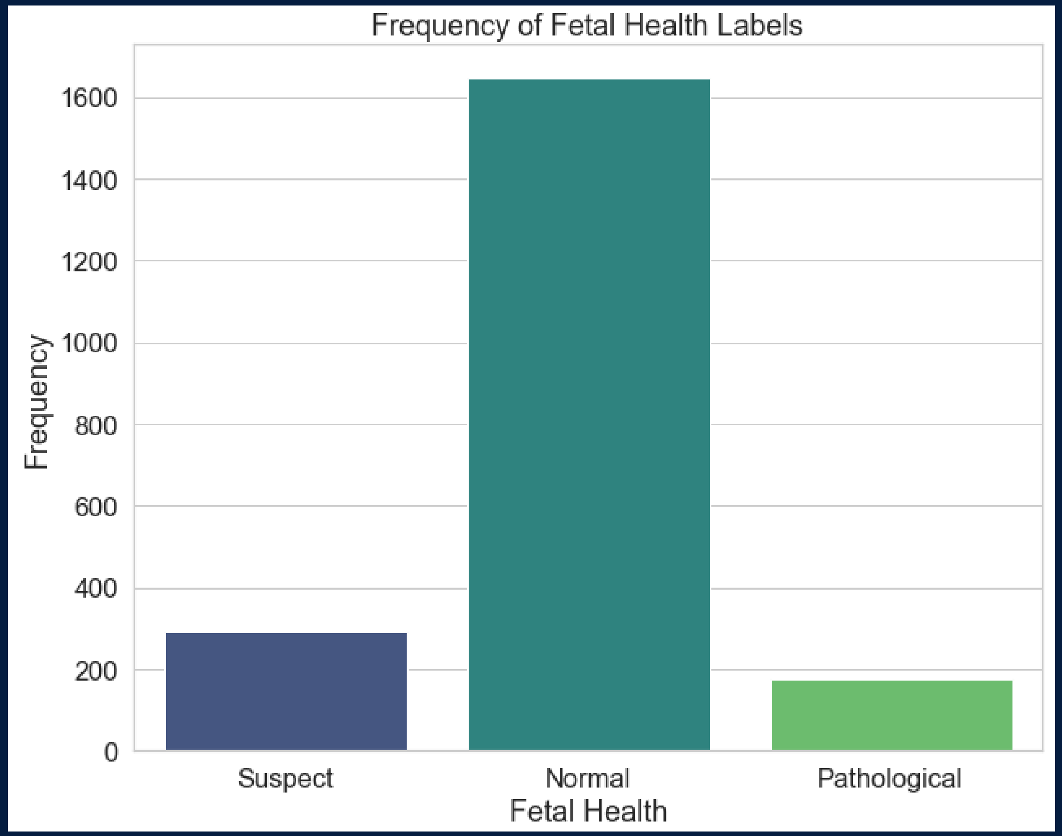


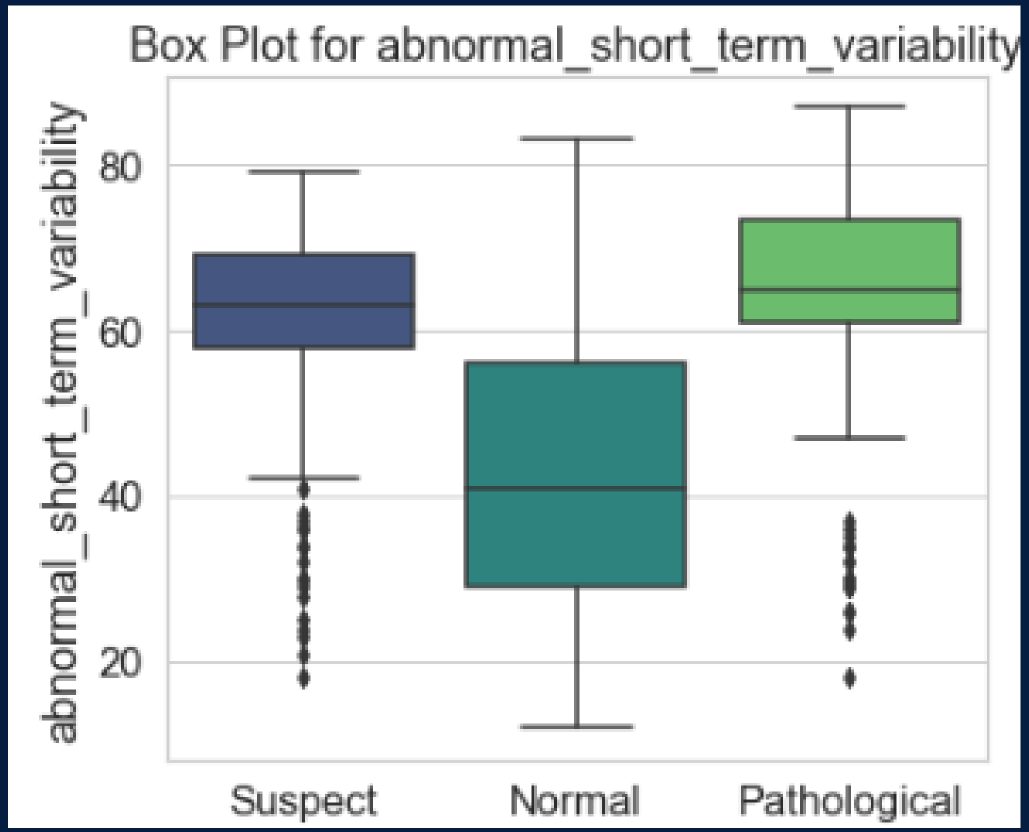
Exploratory Data Analysis

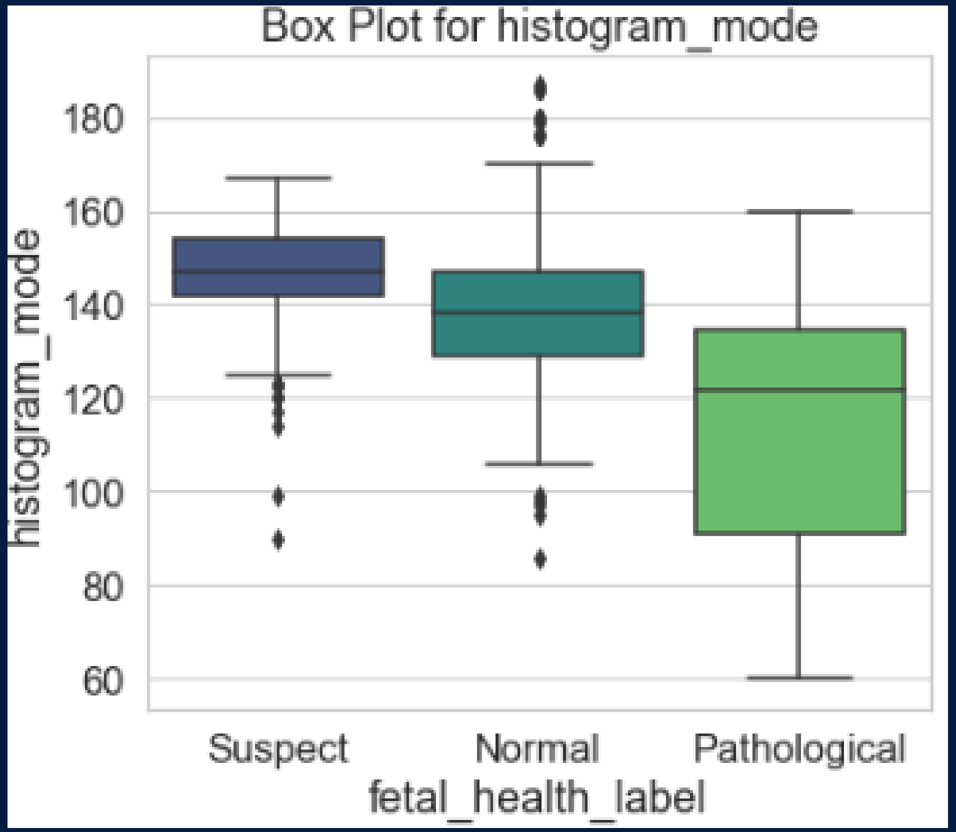
We observed the distribution of the target classes among multiple classes. We found out significant differences among some features, as well as class imbalance in the dataset.





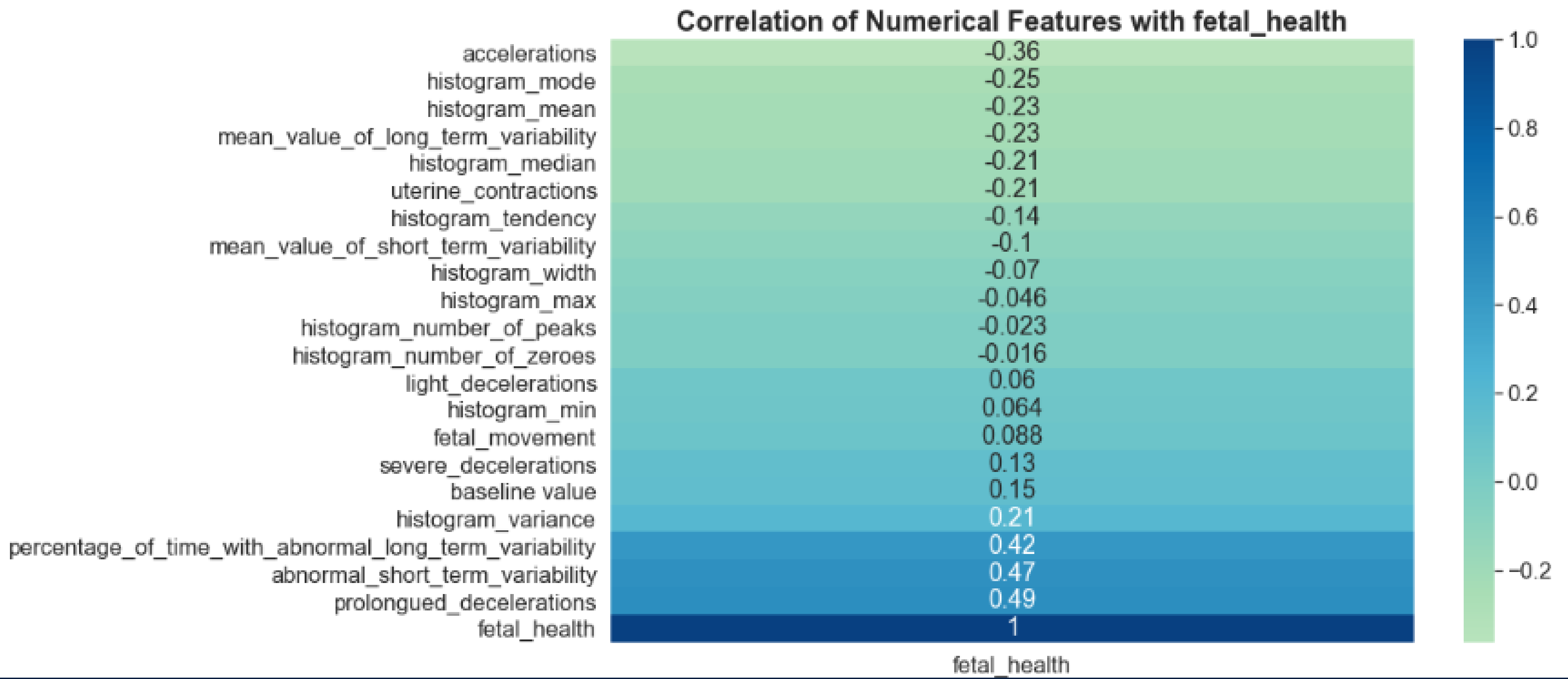






CORRELATIONS WITH TARGET VARIABLE

By looking at the correlation coefficients of the features and the target variable fetal_health, we see that three features have a linear correlation larger than 0.40. This significant correlation underscores the potential impact of these features on predicting fetal health outcomes, highlighting their importance in our machine learning approach for fetal risk classification.



MODEL TRAINING AND EVALUATION

80%
TRAINING

20%
TEST

To optimize our models, hyperparameters are tuned using GridSearchCV. We rigorously evaluate the selected model using ten-fold cross-validation to ensure precise performance evaluation, independent of random data partitions that may influence metrics.

Finally, the best-performing model is saved using the pickle library for future use and deployment.

RESULTS

After running all experiments, it is clear to see that the Neural Network model performs generally better at classifying the data instances.

Model	Balanced Accuracy	F1 Score	Precission
Baseline Model	35.0%	27.0%	81.0%, 16%, and 6% for 'Normal', 'Suspect', and 'Pathological' classes respectively.
Logistic Regression	68.60%	69.75%	Ranged from 60% to 82% for different health classes
Random Forest	71.82%	75.10%	84.46%, 66%, and 87% for 'Normal', 'Suspect', and 'Pathological' classes respectively.
Neural Network	92.11%	85.16%	Demonstrated robust performance in accurately classifying fetal health conditions across different health categories.

CONCLUSIONS

In comparing the performance of the three models, Logistic Regression served as a foundational approach, offering initial insights but demonstrating relatively lower performance metrics. On the other hand, the Random Forest model surpassed Logistic Regression, demonstrating robust performance across various evaluation metrics. However, the Neural Network emerged as the standout performer, showcasing superior accuracy and predictive capability, thus highlighting its effectiveness in accurately classifying fetal health conditions.