Fetal Health Classification from CTG Data

Executive Summary

Monitoring fetal health is essential for ensuring the well-being of both pregnant women and their babies. Cardiotocogram (CTG) is a crucial tool in this process, recording key indicators such as fetal heart rate, movements, and uterine contractions. These data help healthcare professionals identify potential issues early, enabling timely medical intervention to prevent complications.

This project aimed to classify fetal health into three categories—Normal, Suspect, and Pathological—using features extracted from CTG data. We explored and compared various supervised machine learning algorithms to determine the most effective model. To address the challenge of class imbalance, we applied the SMOTE (Synthetic Minority Over-sampling Technique), which improved the dataset's balance and enhanced model performance.

Our analysis revealed the Random Forest classifier as the top-performing model, delivering exceptional accuracy and reliability across all classes. Gradient Boosting also showed strong performance, particularly in distinguishing between classes, while Neural Networks provided balanced precision and recall. These models demonstrated robustness in handling the dataset's complexity.

Key achievements include successfully managing class imbalance with SMOTE, implementing multiple machine learning algorithms, and achieving high classification performance, particularly with tree-based models like Random Forest and Gradient Boosting. These findings highlight the potential of machine learning to improve fetal health assessments, providing reliable tools for healthcare professionals.

Future work will focus on validating the model with data from different sources, experimenting with Deep Learning for automated feature extraction, and enhancing model interpretability to ensure actionable insights for healthcare providers.

Contents

1. Background and Introduction

Background

Fetal health monitoring is a critical aspect of prenatal care, with Cardiotocogram (CTG) data serving as a key tool for assessing fetal well-being. CTG provides a non-invasive method to track the fetus's heart rate, movements, and uterine contractions, offering vital insights into the baby's condition. This monitoring is essential for detecting potential issues early and ensuring timely medical intervention.

The Problem

The central challenge in this project is the accurate classification of fetal health into three distinct categories: Normal, Suspect, and Pathological. This classification is derived from features extracted from CTG data and is crucial for making informed decisions regarding medical care. The project addresses several key questions:

- Can we accurately classify fetal health status using features extracted from CTG data?
- Which machine learning algorithms perform best for this multi-class classification task?

The Goal of Study

The primary goal of this study is to develop a machine learning model capable of accurately classifying fetal health status based on CTG data. By identifying the most effective algorithms, this project aims to enhance the accuracy and reliability of fetal health assessments.

2. Data Preparation and Preprocessing

Data Source

The dataset for this project is publicly available and can be accessed on Kaggle under the title "Fetal Health Classification". The data has been referenced from Ayres de Campos et al. (2000) in their work, "SisPorto 2.0: A Program for Automated Analysis of Cardiotocograms," published in *J Matern Fetal Med*, 5:311-318.

Data Description

The dataset, named 'fetal_health.csv', contains 2,126 records, each with 22 features extracted from CTG exams. These features include various measurements related to fetal heart rate, fetal movements, and uterine contractions. The data has been categorized by three expert obstetricians into three classes: Normal, Suspect, and Pathological. The features and their descriptions are as follows:

- **Baseline value**: Baseline Fetal Heart Rate (FHR)
- **Accelerations**: Number of accelerations per second
- **Fetal movement**: Number of fetal movements per second
- **Uterine contractions**: Number of uterine contractions per second
- **Light decelerations**: Number of light decelerations per second
- **Severe decelerations**: Number of severe decelerations per second
- **Prolonged decelerations**: Number of prolonged decelerations per second
- **Abnormal short-term variability**: Percentage of time with abnormal short-term variability
- **Mean value of short-term variability**
- **Percentage of time with abnormal long-term variability**
- **Mean value of long-term variability**
- **Histogram width**: Width of the histogram constructed from a study
- **Histogram min**: Histogram minimum value
- **Histogram max**: Histogram maximum value

- **Histogram number of peaks**: Number of peaks in the exam histogram
- **Histogram number of zeroes**: Number of zeroes in the exam histogram
- **Histogram mode**: Histogram mode
- **Histogram mean**: Histogram mean
- **Histogram median**: Histogram median
- **Histogram variance**: Histogram variance
- **Histogram tendency**: Histogram trend
- **Fetal health**: 1 - Normal, 2 - Suspect, 3 - Pathological

Data Statistics

Missing value detection revealed that there are no missing values in the dataset, indicating that imputation is unnecessary. The summary statistics show that different variables have varying scales and magnitudes, suggesting that normalization or standardization may be necessary before using these features for modeling. Upon examining the data types of each column, we observe that all features appear to be numeric, except for the target column, which is categorical in nature.

```
Missing values in the dataset:
baseline value                                            0
accelerations                                             0
fetal_movement                                            0
uterine_contractions                                      0
light_decelerations                                       0
severe_decelerations                                      0
prolongued_decelerations                                  0
abnormal_short_term_variability                           0
mean_value_of_short_term_variability                      0
percentage_of_time_with_abnormal_long_term_variability    0
mean_value_of_long_term_variability                       0
histogram_width                                           0
histogram_min                                             0
histogram_max                                             0
histogram_number_of_peaks                                 0
histogram_number_of_zeroes                                0
histogram_mode                                            0
histogram_mean                                            0
histogram_median                                          0
histogram_variance                                        0
histogram_tendency                                        0
fetal_health                                              0
dtype: int64
```

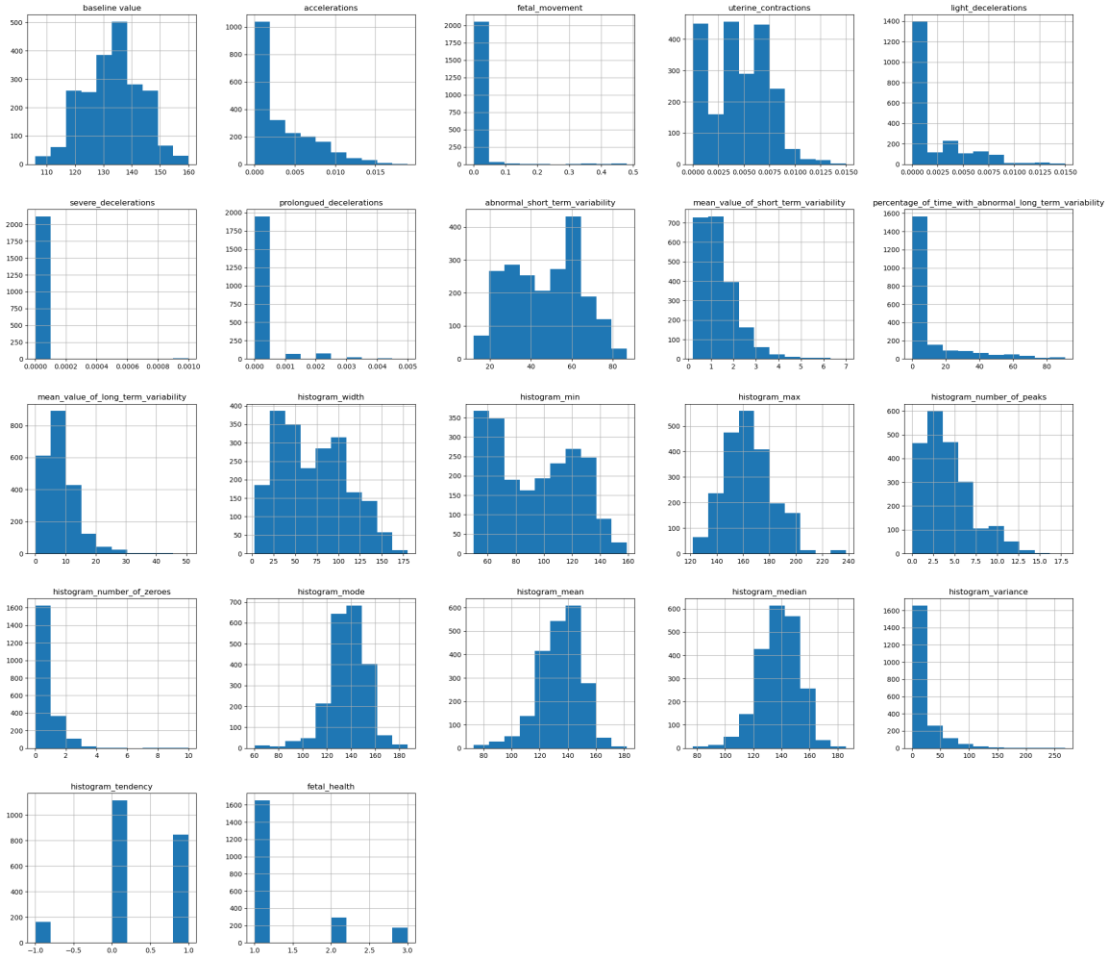
	count	mean	std	min	25%	50%	75%	max
baseline value	2126.0	133.303857	9.840844	106.0	126.000	133.000	140.000	160.000
accelerations	2126.0	0.003178	0.003866	0.0	0.000	0.002	0.006	0.019
fetal_movement	2126.0	0.009481	0.046666	0.0	0.000	0.000	0.003	0.481
uterine_contractions	2126.0	0.004366	0.002946	0.0	0.002	0.004	0.007	0.015
light_decelerations	2126.0	0.001889	0.002960	0.0	0.000	0.000	0.003	0.015
severe_decelerations	2126.0	0.000003	0.000057	0.0	0.000	0.000	0.000	0.001
prolongued_decelerations	2126.0	0.000159	0.000590	0.0	0.000	0.000	0.000	0.005
abnormal_short_term_variability	2126.0	46.990122	17.192814	12.0	32.000	49.000	61.000	87.000
mean_value_of_short_term_variability	2126.0	1.332785	0.883241	0.2	0.700	1.200	1.700	7.000
percentage_of_time_with_abnormal_long_term_variability	2126.0	9.846660	18.396880	0.0	0.000	0.000	11.000	91.000
mean_value_of_long_term_variability	2126.0	8.187629	5.628247	0.0	4.600	7.400	10.800	50.700
histogram_width	2126.0	70.445908	38.955693	3.0	37.000	67.500	100.000	180.000
histogram_min	2126.0	93.579492	29.560212	50.0	67.000	93.000	120.000	159.000
histogram_max	2126.0	164.025400	17.944183	122.0	152.000	162.000	174.000	238.000
histogram_number_of_peaks	2126.0	4.068203	2.949386	0.0	2.000	3.000	6.000	18.000
histogram_number_of_zeroes	2126.0	0.323612	0.706059	0.0	0.000	0.000	0.000	10.000
histogram_mode	2126.0	137.452023	16.381289	60.0	129.000	139.000	148.000	187.000
histogram_mean	2126.0	134.610536	15.593596	73.0	125.000	136.000	145.000	182.000
histogram_median	2126.0	138.090310	14.466589	77.0	129.000	139.000	148.000	186.000
histogram_variance	2126.0	18.808090	28.977636	0.0	2.000	7.000	24.000	269.000
histogram_tendency	2126.0	0.320320	0.610829	-1.0	0.000	0.000	1.000	1.000
fetal_health	2126.0	1.304327	0.614377	1.0	1.000	1.000	1.000	3.000

3. Data Exploration and Visualization

Data Visualization

1) Histogram to view Distributions of features

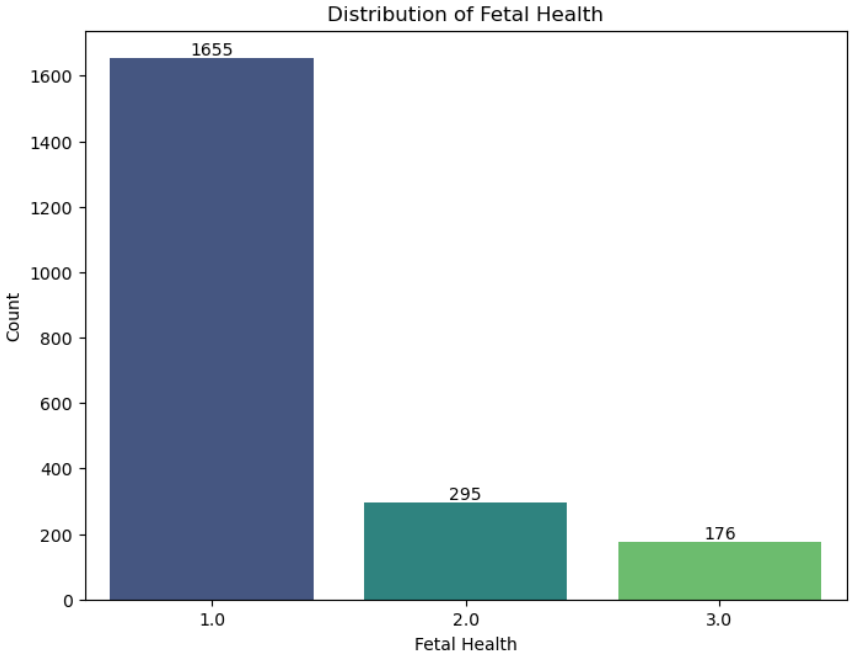
The distribution of various features in the dataset was initially explored using histograms. It was observed that many features, such as accelerations, fetal movement, and uterine contractions, exhibited a highly skewed distribution with most values clustered near zero. In contrast, the baseline fetal heart rate showed a distribution that approximated normality. Features related to variability, such as abnormal short-term variability and mean short-term variability, had broader distributions, indicating a wider range of values across the dataset.



2) Bar chart to show Distribution of Target

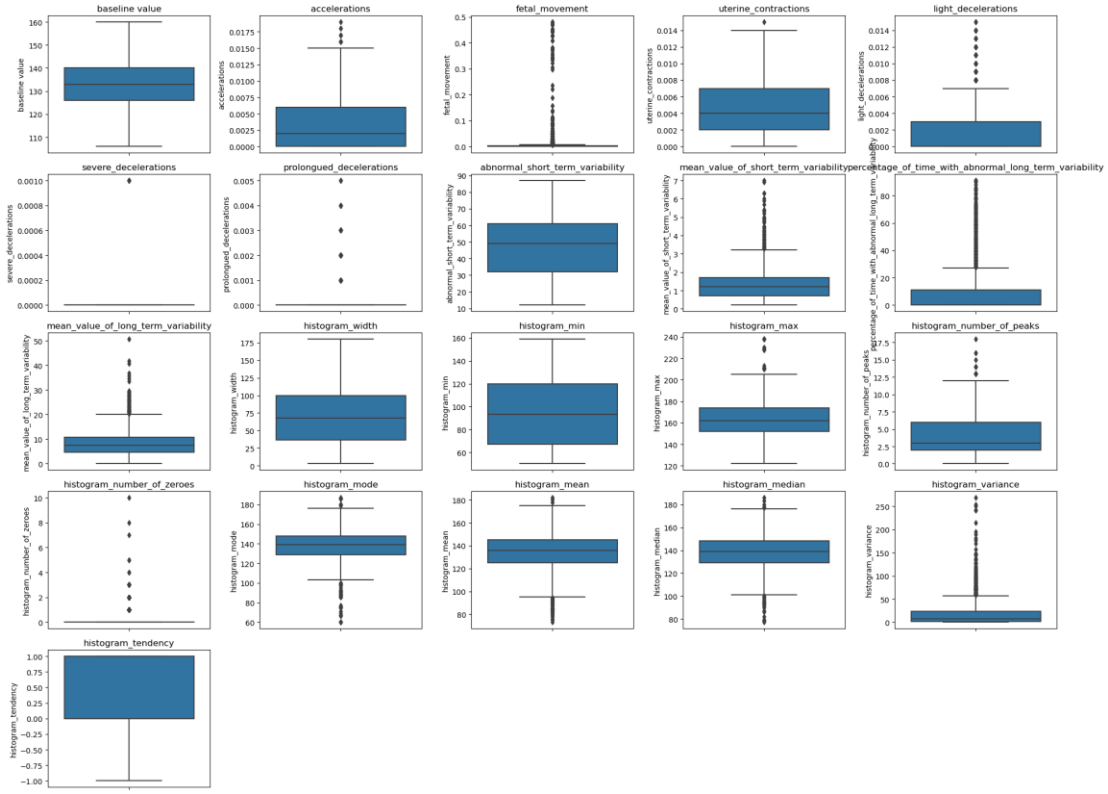
A bar chart was used to examine the distribution of the target variable, revealing an imbalance in the dataset. The majority of the cases were classified as Normal, with fewer instances categorized as Suspect and even fewer as Pathological. This imbalance highlights the need for

careful consideration in model development to ensure that less-represented classes are adequately predicted.



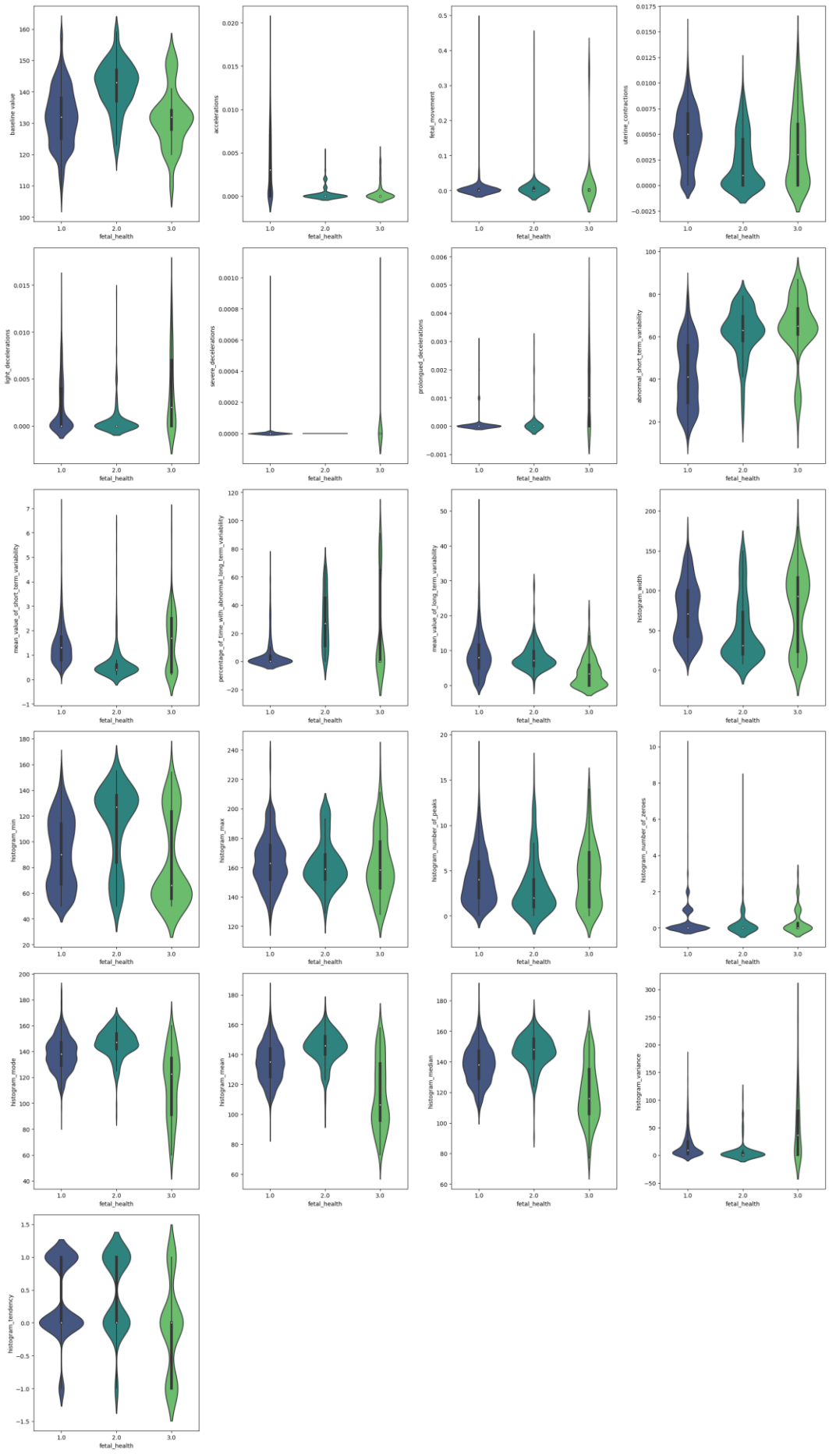
3) Box plots to detect outliers

Box plots were utilized to identify outliers within the dataset. Significant outliers were detected in features such as accelerations, fetal movement, uterine contractions, light decelerations, prolonged decelerations, severe decelerations, and the mean value of long-term variability. Additionally, features like histogram min, histogram max, histogram mean, and histogram median also exhibited outliers, though to a lesser extent. These outliers, derived from CTG reports, were not removed as they likely represent true variations within the population rather than errors, thus retaining them helps in avoiding overfitting the models.



4) Violin plots

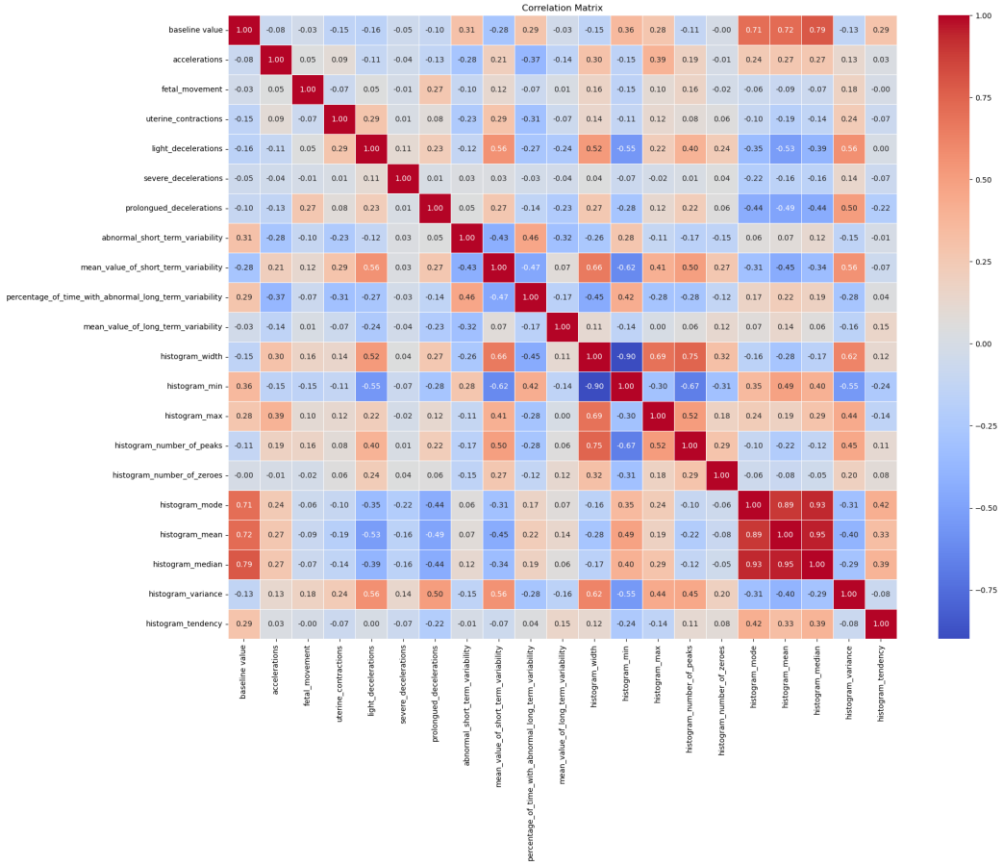
Violin plots provided insights into the distribution patterns of various features across the three fetal health categories. The plots indicated distinct distribution patterns for features such as baseline value, abnormal short-term variability, and mean short-term variability, underscoring their potential importance in distinguishing between different health conditions. Additionally, features like accelerations and uterine contractions showed significant variation, particularly between healthy and unhealthy states, further highlighting their relevance in fetal health assessment.


baseline value
accelerations
fetal_movement
uterine_contractions
light_decelerations
severe_decelerations
prolongued_decelerations
abnormal_short_term_variability
mean_value_of_short_term_variability
percentage_of_time_with_abnormal_long_term_variability
mean_value_of_long_term_variability
histogram_width
histogram_min
histogram_max
histogram_number_of_peaks
histogram_number_of_zeroes
histogram_mode
histogram_mean
histogram_median
histogram_variance
histogram_tendency
fetal_health
1.0
2.0
3.0

5) Correlation matrix

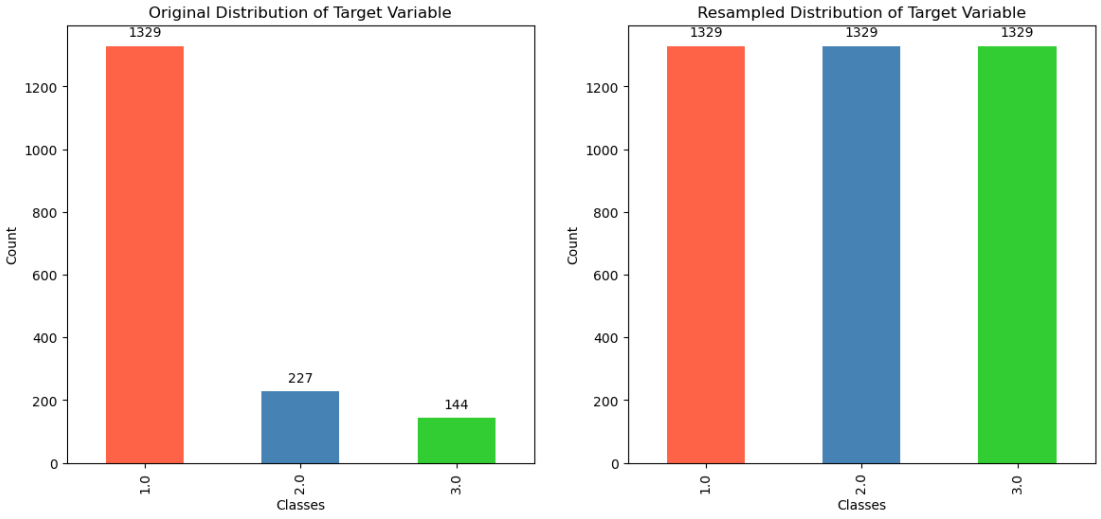
The correlation matrix, visualized through a heatmap, revealed that certain features, such as "histogram_mean," "histogram_median," and "histogram_mode," were highly correlated with one another. This strong correlation suggests redundancy, making these features candidates for removal by setting up threshold as 0.75. The analysis of the target variable in relation to other features revealed that prolonged decelerations and abnormal short-term variability were positively correlated with fetal health issues, while higher accelerations were negatively correlated, indicating better fetal health outcomes.

Based on the correlation matrix, it was decided to drop the highly correlated features—histogram min, histogram mean, and histogram median—to streamline the dataset. The remaining features include baseline value, accelerations, fetal movement, uterine contractions, light decelerations, severe decelerations, prolonged decelerations, abnormal short-term variability, mean value of short-term variability, percentage of time with abnormal long-term variability, mean value of long-term variability, histogram width, histogram max, histogram number of peaks, histogram number of zeroes, histogram mode, histogram variance, histogram tendency, and fetal health. This refined feature set will be used in the subsequent modeling efforts.



4. Data Mining Techniques and Implementation

To address the issue of imbalanced class labels in the dataset, the SMOTE (Synthetic Minority Over-sampling Technique) method was employed. By generating synthetic instances of the minority classes, SMOTE ensured that the dataset had an equal number of instances for each class. This balancing of the labels was crucial to prevent the models from being biased toward the majority class and to enhance overall classification performance.



Following this, the dataset was partitioned into training and validation sets, with 80% of the data allocated for training the models and the remaining 20% reserved for validation. This 80:20 split allowed for sufficient data to train robust models while also providing a separate set of data to evaluate their performance objectively.

Standardization was applied to all feature variables in both the training and validation datasets. This step was necessary to ensure that all features had a consistent scale, as varying scales could adversely affect the performance of certain machine learning algorithms, particularly those that rely on distance metrics.

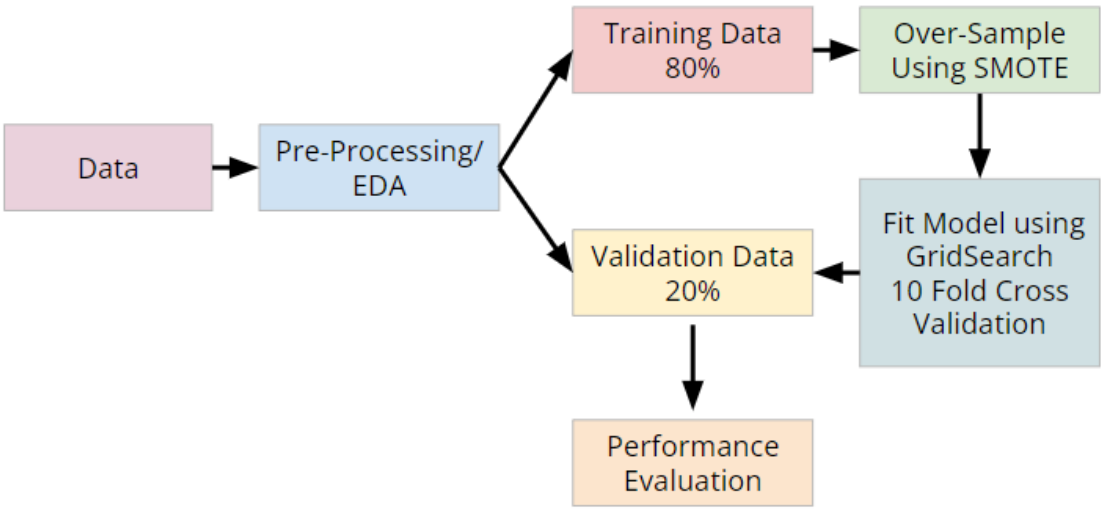
For model selection, GridSearchCV with 10-fold cross-validation was employed. Given our dataset of 3,987 samples in the training set after oversampling, 10-fold cross-validation provided a thorough evaluation without significantly increasing computation time. This method, a widely accepted standard in machine learning, involves dividing the data into 10 subsets, where each subset is used once as a validation set while the remaining nine subsets are used for training. This approach ensures that the model generalizes well to unseen data by balancing bias and variance and by reducing the likelihood of overfitting. GridSearchCV was

utilized to systematically explore various hyperparameter combinations for each model, ultimately selecting the configuration that yielded the highest accuracy.

The following algorithms are applied to address the classification problem:

- KNN
- Gaussian Naive Bayes
- Random Forest
- Gradient Boosting
- Logistic Regression
- Linear Discriminant Analysis
- Linear Discriminant Analysis (trained without outliers)
- Neural Network
- Support Vector Machine

The flow chart below exhibits how models are implemented:



The best parameters for each model were summarized in the table below:

Model	**Best Parameters**
KNN	metric='manhattan', n_neighbors=2, weights='distance'
Naïve Bayes Classifier	var_smoothing=0.0035111
Random Forest	bootstrap=False, max_depth=None, min_samples_leaf=1, min_samples_split=2, n_estimators=50
Gradient Boosting Classifier	learning_rate=0.1, max_depth=4, min_samples_leaf=2, min_samples_split=5, n_estimators=100, subsample=1.0
Logistic Regression	C=100, solver='lbfgs'
Linear Discriminant Analysis	shrinkage=None, solver='svd'
SVM	C=100, gamma=0.1, kernel='rbf'
Neural Network	3 hidden layers of sizes 8, 16, 24

5. Performance Evaluation

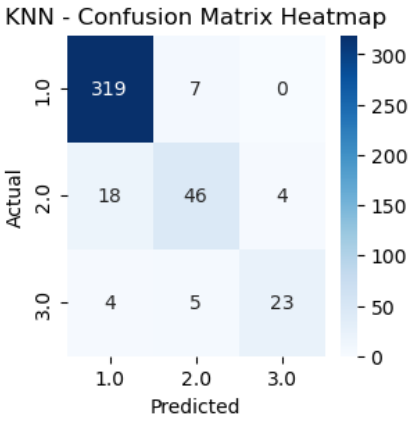
1. Classification Report and Confusion Matrix for each model

1)

```
--- KNN ---
Accuracy: 0.9107981220657277
Classification Report:
              precision    recall  f1-score   support

         1.0       0.94      0.98      0.96       326
         2.0       0.79      0.68      0.73        68
         3.0       0.85      0.72      0.78        32

    accuracy                           0.91       426
   macro avg       0.86      0.79      0.82       426
weighted avg       0.91      0.91      0.91       426
```



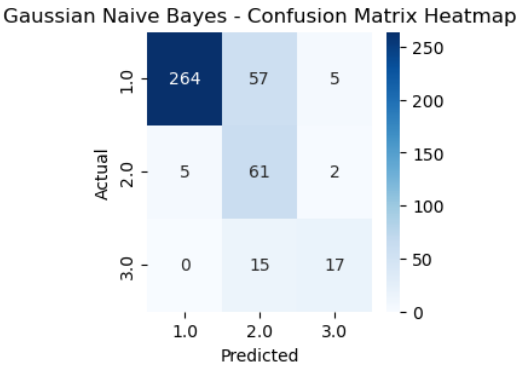
The KNN model achieved a solid accuracy of 0.91, with strong performance in predicting the "Normal" class (Class 1), demonstrated by high precision and recall. However, it struggled with the "Suspect" (Class 2) and "Pathologic" (Class 3) classes, where recall values were notably lower. The confusion matrix reveals that while the model effectively distinguishes the "Normal" class, there is some confusion between the "Suspect" and "Pathologic" classes, indicating a potential area for improvement.

2)

```
--- Gaussian Naive Bayes ---
Accuracy: 0.8028169014084507
Classification Report:
              precision    recall  f1-score   support

         1.0       0.98      0.81      0.89       326
         2.0       0.46      0.90      0.61        68
         3.0       0.71      0.53      0.61        32

    accuracy                           0.80       426
   macro avg       0.72      0.75      0.70       426
weighted avg       0.88      0.80      0.82       426
```



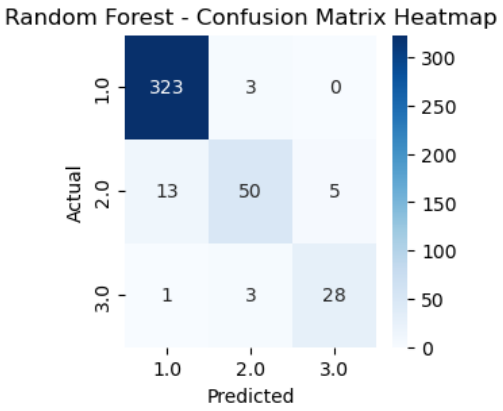
The Gaussian Naive Bayes model showed an accuracy of 0.80, with strong precision (0.98) but lower recall (0.81) for identifying Class 1 cases. It was less reliable in predicting Class 2 and Class 3 cases, with low precision for Class 2 (0.46) and moderate precision for Class 3 (0.71). The confusion matrix indicates significant misclassification between "normal" and "suspect" cases and poor performance in distinguishing "pathological" cases.

3)

```
--- Random Forest ---
Accuracy: 0.9413145539906104
Classification Report:
              precision    recall  f1-score   support

         1.0       0.96      0.99      0.97       326
         2.0       0.89      0.74      0.81        68
         3.0       0.85      0.88      0.86        32

    accuracy                           0.94       426
   macro avg       0.90      0.87      0.88       426
weighted avg       0.94      0.94      0.94       426
```



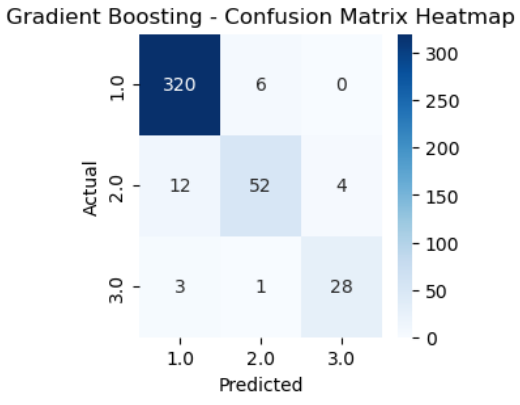
The Random Forest model performed the best among the models, with an accuracy of 0.94. It demonstrated excellent precision and recall for Class 1, Class 2, and Class 3 cases, particularly excelling in identifying "normal" cases with nearly perfect recall (0.99). The confusion matrix shows minimal misclassification, making Random Forest a reliable model for distinguishing between all three classes with high accuracy.

4)

```
--- Gradient Boosting ---
Accuracy: 0.9389671361502347
Classification Report:
              precision    recall  f1-score   support

         1.0       0.96      0.98      0.97       326
         2.0       0.88      0.76      0.82        68
         3.0       0.88      0.88      0.88        32

    accuracy                           0.94       426
   macro avg       0.90      0.87      0.89       426
weighted avg       0.94      0.94      0.94       426
```



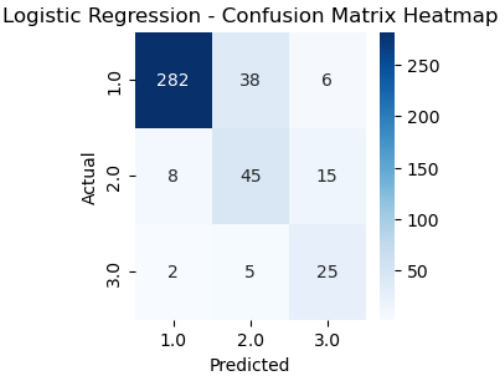
The Gradient Boosting model also showed strong performance with an accuracy of 0.94, similar to the Random Forest model. It effectively identified Class 1, Class 2, and Class 3 cases, with balanced precision and recall across all classes. The confusion matrix indicates minimal confusion, particularly excelling in correctly identifying "pathological" cases, which had a recall of 0.88.

5)

```
--- Logistic Regression ---
Accuracy: 0.8262910798122066
Classification Report:
              precision    recall  f1-score   support

         1.0       0.97      0.87      0.91       326
         2.0       0.51      0.66      0.58        68
         3.0       0.54      0.78      0.64        32

    accuracy                           0.83       426
   macro avg       0.67      0.77      0.71       426
weighted avg       0.86      0.83      0.84       426
```



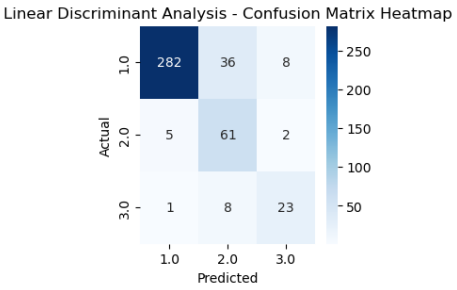
The Logistic Regression model achieved an accuracy of 0.83, performing well in predicting Class 1 with high precision (0.97) but showing weaknesses in identifying Class 2and Class 3 cases, with lower precision values of 0.51 and 0.54, respectively. The confusion matrix reflects significant misclassification, especially between "normal" and "suspect" cases, and between "suspect" and "pathological" cases, indicating this model is less effective at distinguishing between these classes.

6)

```
--- Linear Discriminant Analysis ---
Accuracy: 0.8591549295774648
Classification Report:
              precision    recall  f1-score   support

         1.0       0.98      0.87      0.92       326
         2.0       0.58      0.90      0.71        68
         3.0       0.70      0.72      0.71        32

    accuracy                           0.86       426
   macro avg       0.75      0.83      0.78       426
weighted avg       0.89      0.86      0.87       426
```



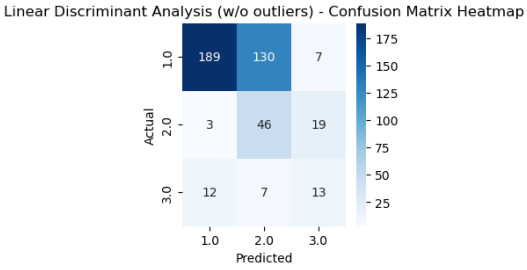
The Logistic Regression model achieved an accuracy of 0.83, performing well in predicting Class 1 with high precision (0.97) but showing weaknesses in identifying Class 2and Class 3 cases, with lower precision values of 0.51 and 0.54, respectively. The confusion matrix reflects significant misclassification, especially between "normal" and "suspect" cases, and between "suspect" and "pathological" cases, indicating this model is less effective at distinguishing between these classes.

7)

```
--- Linear Discriminant Analysis (w/o outliers) ---
Accuracy: 0.5821596244131455
Classification Report:
              precision    recall  f1-score   support

         1.0       0.93      0.58      0.71       326
         2.0       0.25      0.68      0.37        68
         3.0       0.33      0.41      0.37        32

    accuracy                           0.58       426
   macro avg       0.50      0.55      0.48       426
weighted avg       0.77      0.58      0.63       426
```



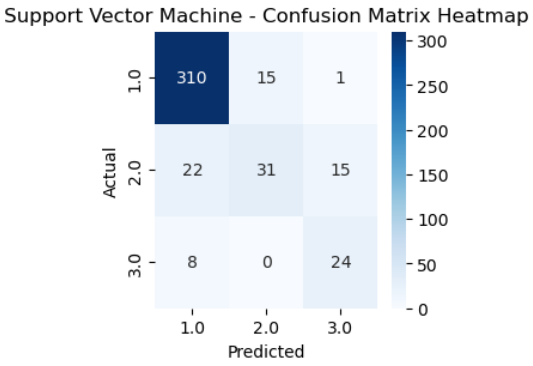
When the outliers were removed in the training dataset but “remains” in the validation dataset , the LDA model’s performance significantly declined, with accuracy dropping to 0.58. The model struggled particularly with Class 2 and Class 3 cases, showing low precision and recall across these classes. The confusion matrix reflects substantial misclassification, particularly between Class 1 and Class 2, suggesting that outliers play a significant role in the model’s ability to differentiate between these classes.

8)

```
--- Support Vector Machine ---
Accuracy: 0.8568075117370892
Classification Report:
              precision    recall  f1-score   support

         1.0       0.91      0.95      0.93       326
         2.0       0.67      0.46      0.54        68
         3.0       0.60      0.75      0.67        32

    accuracy                           0.86       426
   macro avg       0.73      0.72      0.71       426
weighted avg       0.85      0.86      0.85       426
```



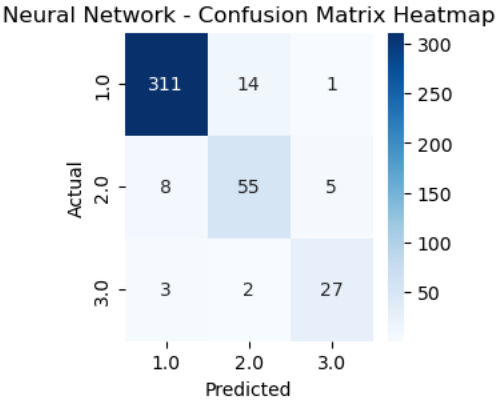
The SVM model also achieved an accuracy of 0.86, performing well in predicting Class 1 with high precision (0.91) and recall (0.95). However, it showed weaker performance for Class 2 and Class 3 cases, with lower precision and recall, particularly for Class 2. The confusion matrix shows significant misclassification between Class 2 and the other two classes, indicating that the SVM model struggles to distinguish "suspect" cases effectively.

9)

```
--- Neural Network ---
Accuracy: 0.9225352112676056
Classification Report:
              precision    recall  f1-score   support

         1.0       0.97      0.95      0.96       326
         2.0       0.77      0.81      0.79        68
         3.0       0.82      0.84      0.83        32

    accuracy                           0.92       426
   macro avg       0.85      0.87      0.86       426
weighted avg       0.92      0.92      0.92       426
```



The Neural Network model exhibited the highest accuracy among these models at 0.92. It demonstrated strong precision and recall across all three classes, particularly excelling in identifying Class 1 cases with precision (0.97) and recall (0.95). For Class 2and Class 3 cases, the model maintained high precision and recall, with values around 0.77 and 0.81 for Class 2, and 0.82 and 0.84 for Class 3. The confusion matrix shows minimal misclassification, indicating that the Neural Network is a robust model for distinguishing between "normal," "suspect," and "pathological" cases.

2. Summary of performance metrics

Model Metrics Sorted by Accuracy:

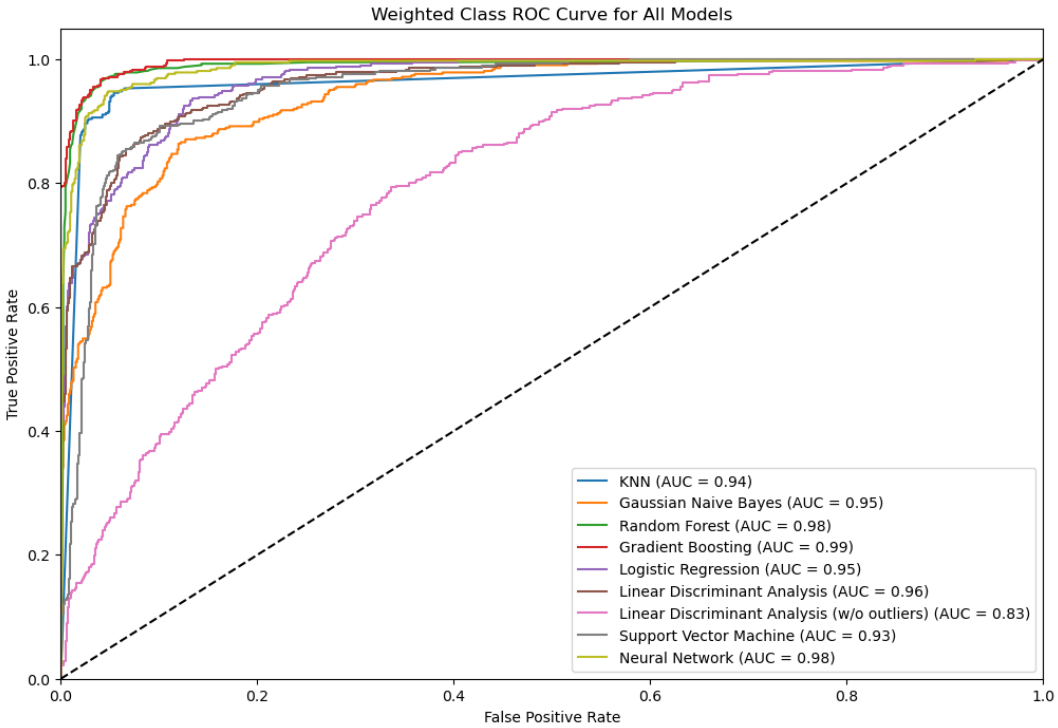
	Accuracy	Best Cross-Validation Accuracy	Weighted Precision	Weighted Recall	Weighted F1-Score	Weighted AUC
Random Forest	0.941315	0.982699	0.939725	0.941315	0.939082	0.980526
Gradient Boosting	0.938967	0.975175	0.937407	0.938967	0.937390	0.988255
Neural Network	0.922535	0.968145	0.924228	0.922535	0.923280	0.977528
KNN	0.910798	0.980187	0.906474	0.910798	0.907103	0.944910
Linear Discriminant Analysis	0.859155	0.827695	0.894404	0.859155	0.868668	0.957914
Support Vector Machine	0.856808	0.982945	0.850379	0.856808	0.849294	0.931397
Logistic Regression	0.826291	0.885378	0.861502	0.826291	0.838634	0.950112
Gaussian Naive Bayes	0.802817	0.773781	0.877453	0.802817	0.821580	0.948808
Linear Discriminant Analysis (w/o outliers)	0.582160	0.863317	0.774153	0.582160	0.631803	0.827278

From the summary table, Random Forest emerges as the top-performing model, with an accuracy of 0.9413 and consistently high metrics across all categories, indicating its robustness in predicting all classes. Gradient Boosting closely follows, with an accuracy of 0.9389 and slightly higher AUC, making it another reliable choice. Neural Network ranks third, offering strong and balanced performance, with a high F1-score and AUC, demonstrating its effectiveness in class prediction.

KNN, LDA, and SVM show moderate performance, with accuracies between 0.9108 and 0.8568. While these models are reasonably effective, they fall short of the top three, particularly in handling class imbalances. Logistic Regression and Gaussian Naive Bayes are less effective, with lower accuracies and weaker performance in precision and recall, making them less reliable for this dataset.

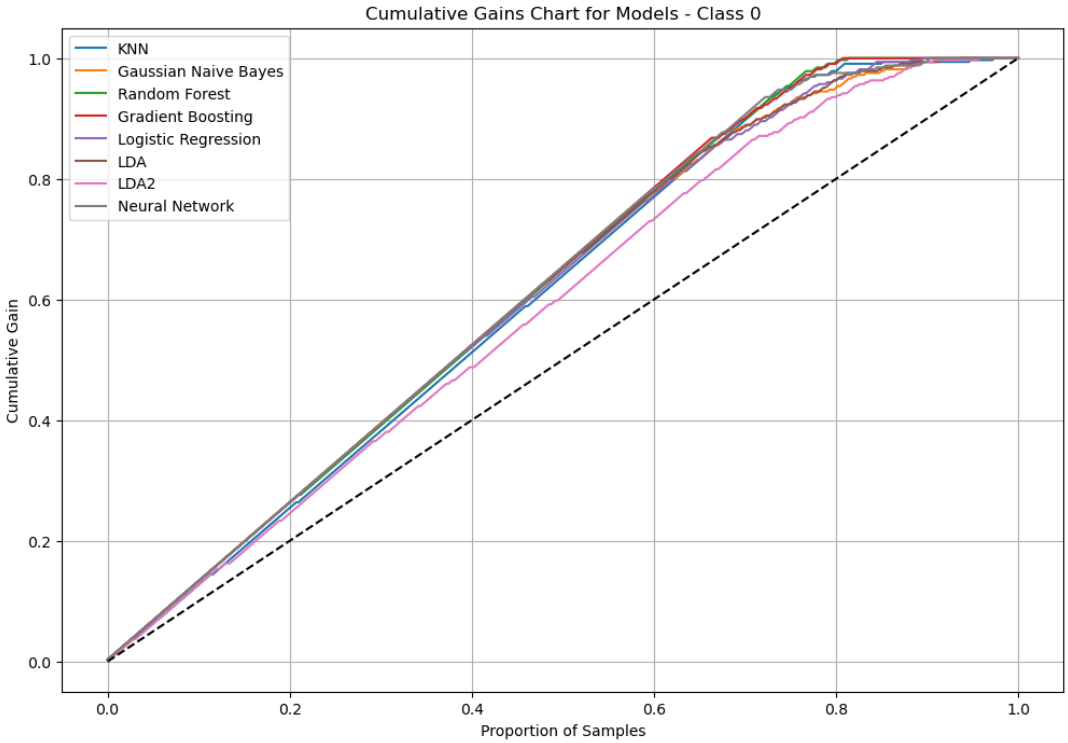
Finally, LDA without outliers shows the poorest performance, with an accuracy of 0.5822, highlighting the critical role outliers play in the model's predictive power. This emphasizes the need for careful handling of outliers to maintain model accuracy.

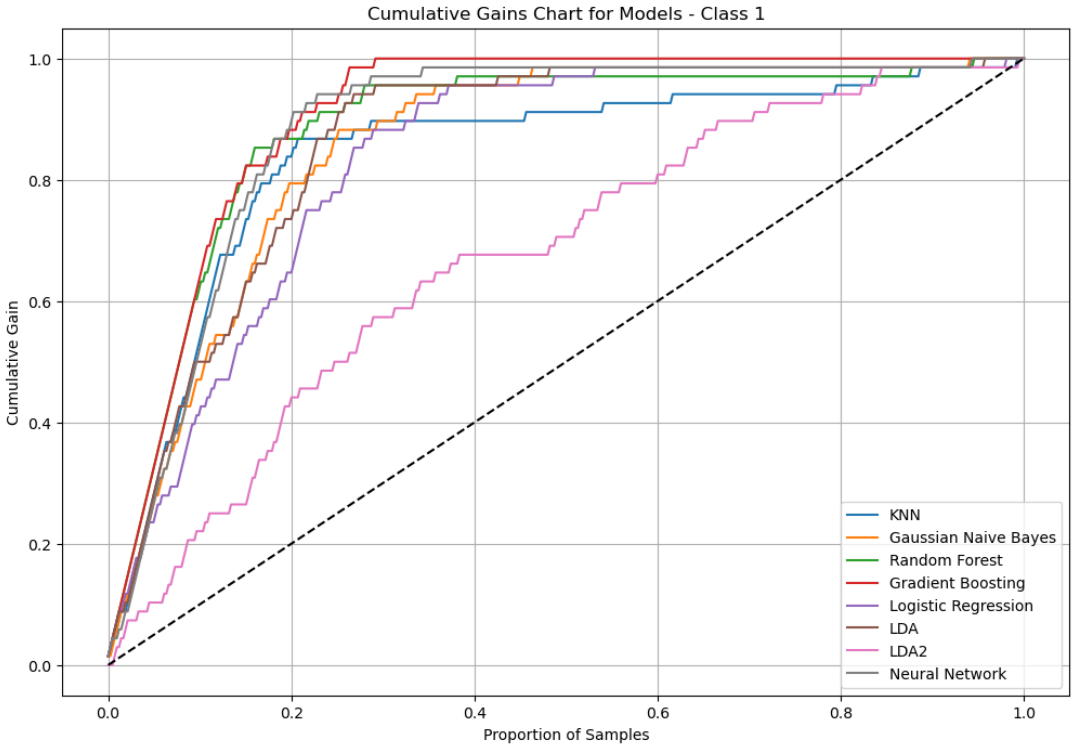
3. Weighted Class ROC Curve

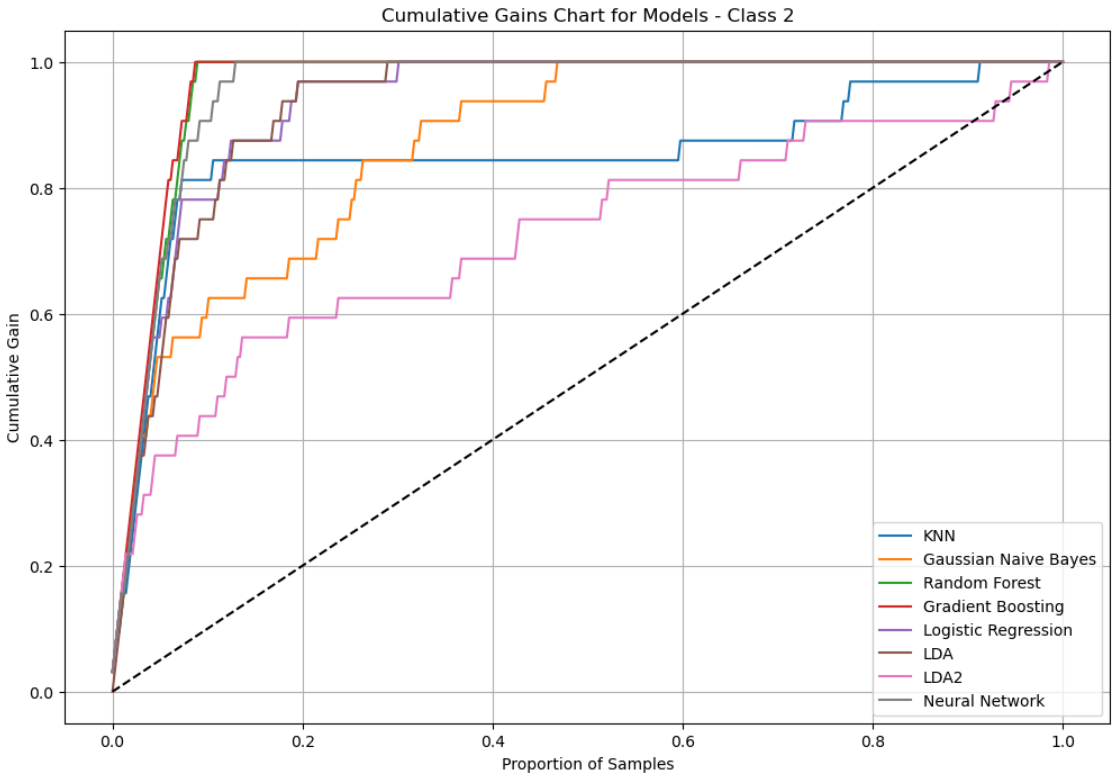


The ROC curve analysis shows a different perspective from the metrics table by focusing on a model's ability to distinguish between classes across various thresholds (AUC). Gradient Boosting stands out with the highest AUC of 0.99, which may not directly correlate with its slightly lower accuracy compared to Random Forest in the metrics table. This indicates that Gradient Boosting is exceptionally good at distinguishing between classes across all possible thresholds, even though Random Forest had the highest accuracy in the metrics table. Gaussian Naive Bayes and Logistic Regression also show strong AUCs (0.95), suggesting good class distinction despite their lower metrics. The differences highlight that AUC provides a broader evaluation of model performance beyond fixed threshold metrics like accuracy and F1-score.

4. Gains Chart







The cumulative gains charts highlight the strong performance of Gradient Boosting and Random Forest across all classes. Both models consistently achieve high cumulative gains early on, indicating their effectiveness in accurately predicting a large proportion of positive cases with fewer samples. This performance is particularly notable in the more challenging Class 1 and Class 2, where these models sharply distinguish themselves from others.

Neural Network also shows solid performance, closely trailing Gradient Boosting and Random Forest. Its gain curves are similarly steep, reflecting its strong predictive power, though it slightly underperforms compared to the top two models.

In contrast, LDA without outliers consistently lags behind, especially in Classes 1 and 2, where its gain curves are the least steep. This underperformance suggests that removing outliers significantly hampers its ability to predict these classes effectively. Other models, like KNN and Gaussian Naive Bayes, show moderate performance, with gain curves that are closer to the diagonal line, indicating less efficient predictions compared to the leading models.

6. Discussion and Recommendation

The results of our analysis underscore the success of various machine learning models in accurately classifying fetal health based on Cardiotocogram (CTG) data. Among the models tested, the Random Forest classifier stood out as the top performer, delivering exceptional accuracy and reliability across all three classes—Normal, Suspect, and Pathological. This model's robustness and ability to handle complex data interactions make it highly suitable for this classification task. Gradient Boosting also demonstrated strong performance, with a slightly lower accuracy but the highest AUC, indicating its superior ability to distinguish between classes across different thresholds.

Neural Networks, while slightly behind Random Forest and Gradient Boosting, proved effective in maintaining balanced precision and recall across all classes, making them a viable option when dealing with datasets requiring the modeling of non-linear relationships.

Models like KNN, LDA, and SVM showed moderate performance, indicating their potential but also their limitations in handling class imbalances and complex data structures. Logistic Regression and Gaussian Naive Bayes were less effective, with lower accuracies and weaker metrics, making them less reliable for this specific task.

A notable observation was the significant decline in performance when outliers were removed in the LDA model, highlighting the importance of retaining these data points. The presence of outliers seems to be integral to capturing meaningful variations in fetal health conditions, rather than being mere noise.

Key Achievements:

- Successfully addressed class imbalance through the application of SMOTE oversampling.
- Implemented and compared multiple machine learning algorithms, identifying Random Forest and Gradient Boosting as the best-performing models.
- Achieved high classification performance, particularly with tree-based models like Random Forest and Gradient Boosting, which excelled in handling the complexity of the data.

Recommendations: Based on the findings, the Random Forest model is recommended as the most effective for this classification task due to its high accuracy and robust performance. Gradient Boosting is also recommended for scenarios where distinguishing between classes across various thresholds is crucial. Neural Networks should be considered in cases where complex, non-linear data relationships need to be modeled. Additionally, retaining outliers in the dataset is crucial, as they contribute significantly to the model’s predictive accuracy.

7. Summary

This study explored and evaluated various machine learning models to classify fetal health into Normal, Suspect, and Pathological categories using features extracted from Cardiotocogram (CTG) data. The Random Forest model emerged as the most effective, achieving the highest accuracy and demonstrating consistent performance across all metrics. Gradient Boosting and Neural Networks also performed well, making them strong alternatives depending on the specific application needs.

Key achievements of this study include overcoming challenges related to class imbalance through SMOTE oversampling, implementing and comparing multiple machine learning algorithms, and achieving high classification performance, particularly with tree-based models. The importance of retaining outliers was also highlighted, as their presence significantly impacted model accuracy.

Future Work:

- Test the model on data collected from other sources to validate its performance across different populations and conditions.
- Experiment with Deep Learning architectures to explore automated feature extraction, which may further enhance classification performance.
- Enhance model interpretability by developing methods to better explain feature importance, ensuring that healthcare professionals can confidently act on the insights provided by the models.

Overall, this study demonstrates the potential of machine learning to improve fetal health assessments, providing healthcare professionals with reliable tools to support decision-making and prevent complications during pregnancy.