Final Project Report
Fetal Health Classification

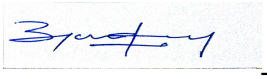
Group 02

Bhavya Pandey
Arya Lokesh Gowda

pandey.bh@northeastern.edu
lokeshgowda.a@northeastern.edu

Percentage of Effort Contributed by Student 1: _____50%_____

Percentage of Effort Contributed by Student 2: _____50%_____

Signature of Student 1: _____  **_____**

Signature of Student 2: ____  **_____**

Submission Date: _____04/12/2024_____

PROJECT SELECTION AND PROBLEM DEFINITION

Problem Setting: Fetal health monitoring is a critical aspect of prenatal care, offering essential insights for healthcare practitioners to ensure optimal outcomes during pregnancy. Accurate assessment of fetal well-being is paramount, as deviations from normal conditions can indicate potential risks and complications. Cardiotocography (CTG) serves as a fundamental tool in this regard, providing a real-time evaluation of fetal heart rate and uterine contractions. By offering continuous monitoring, CTG enables practitioners to detect early signs of fetal distress, guiding timely interventions and facilitating informed decision-making. The comprehensive understanding of fetal health through CTG interpretation enhances the ability of healthcare professionals to safeguard the well-being of both the mother and the unborn child throughout the pregnancy journey.

Problem Definition: This project aims to develop a robust fetal health classification system using machine learning algorithms in Python. Focused on Cardiotocography (CTG), a vital tool in monitoring fetal heart rate and uterine contractions during pregnancy, the objective is to create a predictive model. By leveraging advanced machine learning algorithms, the model will classify CTG outcomes, aiding in the early detection of fetal distress and providing valuable insights for healthcare practitioners. The ultimate goal is to contribute to improved fetal well-being by offering an efficient and accurate classification system for CTG results, thereby assisting in timely interventions and risk assessment during pregnancy.

DATA COLLECTION

Data Source: The CTG dataset has been taken from UC Irvine's Machine Learning Repository, (https://archive.ics.uci.edu/dataset/193/cardiotocography). The original study can be found at https://onlinelibrary.wiley.com/doi/10.1002/1520-6661(200009/10)9:5%3C311::AID-MFM12%3E3.0.CO;2-9.

Data Description: This dataset comprises a total of 2126 instances. There are a total of 21 attributes and 1 target attribute - `fetal_health`, which indicates if the instance is a Normal (1), Suspect (2) or a Pathological (3) class. The attributes have been sub-categorized into the following six groups -

Fetal Heart Rate: Baseline FHR (beats per minute)**,** Number of accelerations, fetal movements, and uterine contractions per second

Decelerations: Light, severe, and prolonged decelerations per second

Short-Term Variability**:** Percentage and mean value of abnormal short-term variability

Long-Term Variability: Percentage and mean value of abnormal long-term variability

Histogram Characteristics: Width, min, max, peaks, zeros, mode, mean, median, variance, tendency

Fetal Health**:** Encoded as 1-Normal; 2-Suspect; 3-Pathological.

Data Dictionary:

1. **Baseline Value (baseline_fetal_heart_rate):**
    - Definition: The baseline Fetal Heart Rate (FHR) measures the average heart rate of the fetus over a specific duration, usually expressed in beats per minute (bpm).
2. **Accelerations (accelerations):**
    - Definition: The number of accelerations per second indicates the occurrences of rapid increases in the fetal heart rate.
3. **Fetal Movement (fetal_movement):**
    - Definition: The number of fetal movements per second represents the instances of movement by the fetus.
4. **Uterine Contractions (uterine_contractions):**
    - Definition: The number of uterine contractions per second reflects the frequency of contractions in the uterus.
5. **Light Decelerations (light_decelerations):**
    - Definition: The number of light decelerations (LDs) per second signifies the occurrences of mild decreases in the fetal heart rate.
6. **Severe Decelerations (severe_decelerations):**
    - Definition: The number of severe decelerations (SDs) per second indicates the instances of significant and abrupt decreases in the fetal heart rate.
7. **Prolonged Decelerations (prolongued_decelerations):**
    - Definition: The number of prolonged decelerations (PDs) per second represents extended periods of decreased fetal heart rate.
8. **Abnormal Short-Term Variability (abnormal_short_term_variability):**
    - Definition: The percentage of time with abnormal short-term variability measures the deviation from the expected short-term variability in the fetal heart rate.
9. **Mean Value of Short-Term Variability (mean_value_of_short_term_variability):**
    - Definition: The mean value of short-term variability represents the average variability in the fetal heart rate over a specific duration.
10. **Percentage of Time with Abnormal Long-Term Variability (percentage_of_time_with_abnormal_long_term_variability):**
    - Definition: This percentage indicates the deviation from the expected long-term variability in the fetal heart rate.
11. **Mean Value of Long-Term Variability (mean_value_of_long_term_variability):**
    - Definition: The mean value of long-term variability represents the average variability in the fetal heart rate over a more extended period.
12. **Histogram Width (histogram_width):**
    - Definition: The width of the histogram made using all values from a CTG record.
13. **Histogram Minimum (histogram_min):**
    - Definition: The minimum value in the histogram created using CTG data.
14. **Histogram Maximum (histogram_max):**

- Definition: The maximum value in the histogram created using CTG data.

15. **Histogram Number of Peaks (histogram_number_of_peaks):**
    - Definition: The number of peaks in the histogram indicates the count of high points in the distribution of CTG data.
16. **Histogram Number of Zeroes (histogram_number_of_zeroes):**
    - Definition: The number of zeros in the histogram represents the count of occurrences where CTG values are zero.
17. **Histogram Mode (histogram_mode):**
    - Definition: The mode of the histogram is the value that appears most frequently.
18. **Histogram Mean (histogram_mean):**
    - Definition: The mean of the histogram represents the average value of CTG data.
19. **Histogram Median (histogram_median):**
    - Definition: The median of the histogram is the middle value when the data is sorted.
20. **Histogram Variance (histogram_variance):**
    - Definition: The variance of the histogram measures the spread or dispersion of CTG data.
21. **Histogram Tendency (histogram_tendency):**
    - Definition: The tendency of the histogram indicates the direction of the distribution, whether it is increasing, decreasing, or constant.

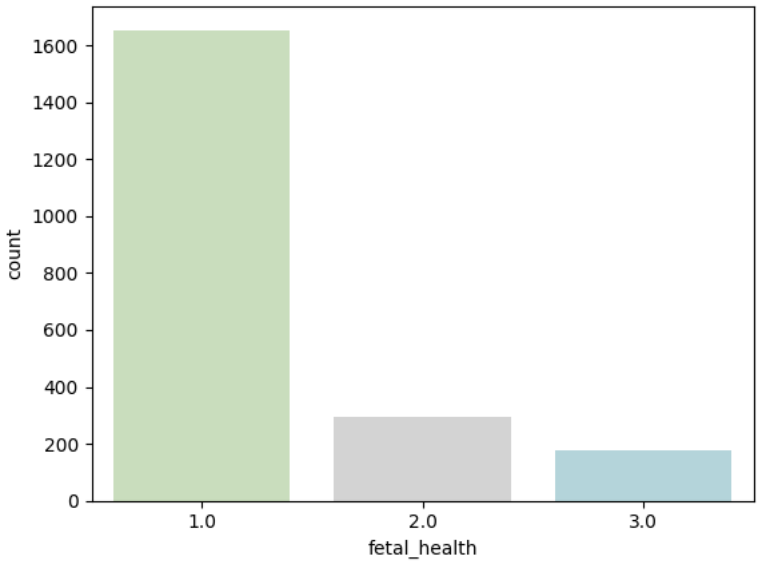
DATA EXPLORATION, VISUALIZATION, AND PROCESSING

Data Exploration

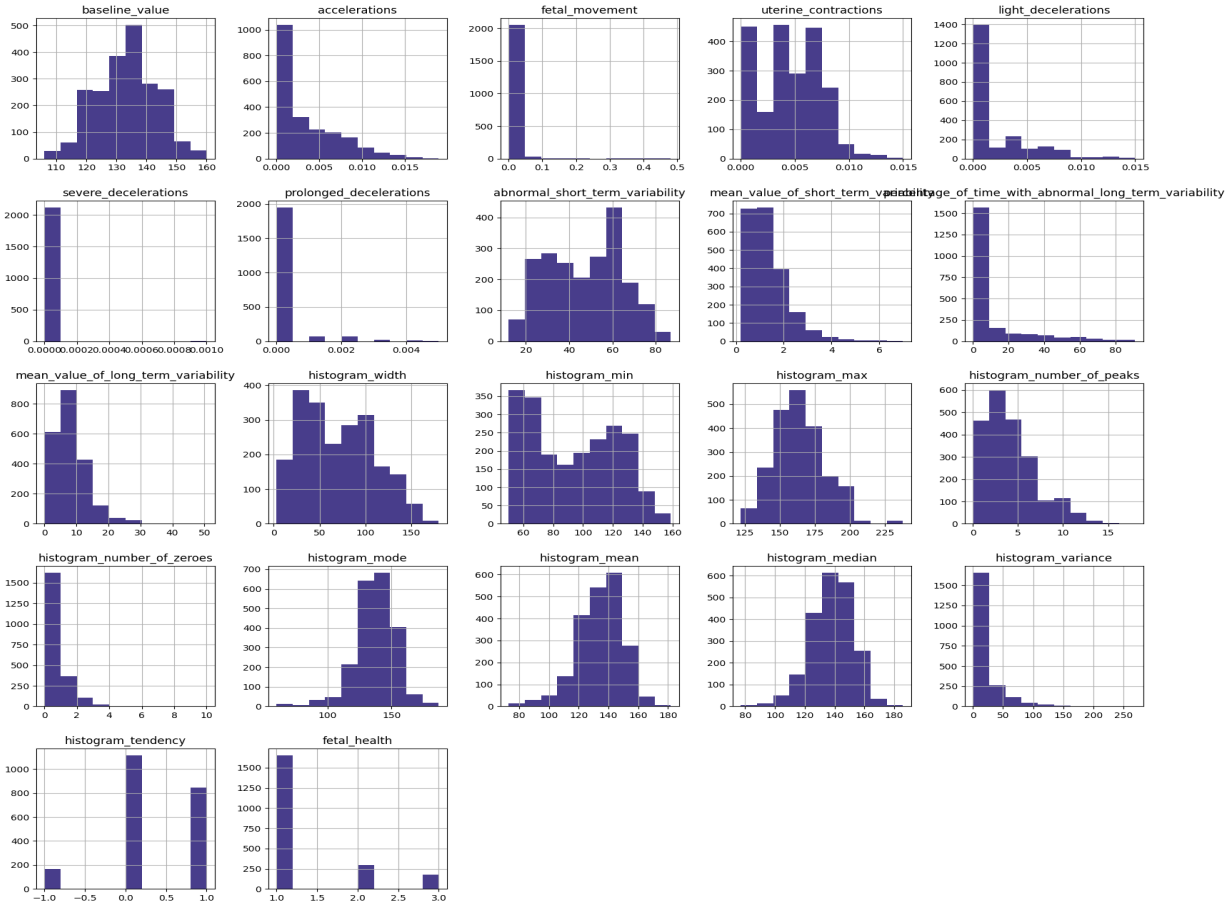
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2126 entries, 0 to 2125
Data columns (total 22 columns):
 #   Column                                                  Non-Null Count  Dtype  
---  ------                                                  --------------  -----  
 0   baseline value                                          2126 non-null   float64
 1   accelerations                                           2126 non-null   float64
 2   fetal_movement                                          2126 non-null   float64
 3   uterine_contractions                                    2126 non-null   float64
 4   light_decelerations                                     2126 non-null   float64
 5   severe_decelerations                                    2126 non-null   float64
 6   prolongued_decelerations                                2126 non-null   float64
 7   abnormal_short_term_variability                         2126 non-null   float64
 8   mean_value_of_short_term_variability                    2126 non-null   float64
 9   percentage_of_time_with_abnormal_long_term_variability  2126 non-null   float64
 10  mean_value_of_long_term_variability                     2126 non-null   float64
 11  histogram_width                                         2126 non-null   float64
 12  histogram_min                                           2126 non-null   float64
 13  histogram_max                                           2126 non-null   float64
 14  histogram_number_of_peaks                               2126 non-null   float64
 15  histogram_number_of_zeroes                              2126 non-null   float64
 16  histogram_mode                                          2126 non-null   float64
 17  histogram_mean                                          2126 non-null   float64
 18  histogram_median                                        2126 non-null   float64
 19  histogram_variance                                      2126 non-null   float64
 20  histogram_tendency                                      2126 non-null   float64
 21  fetal_health                                            2126 non-null   float64
dtypes: float64(22)
memory usage: 365.5 KB
```

	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

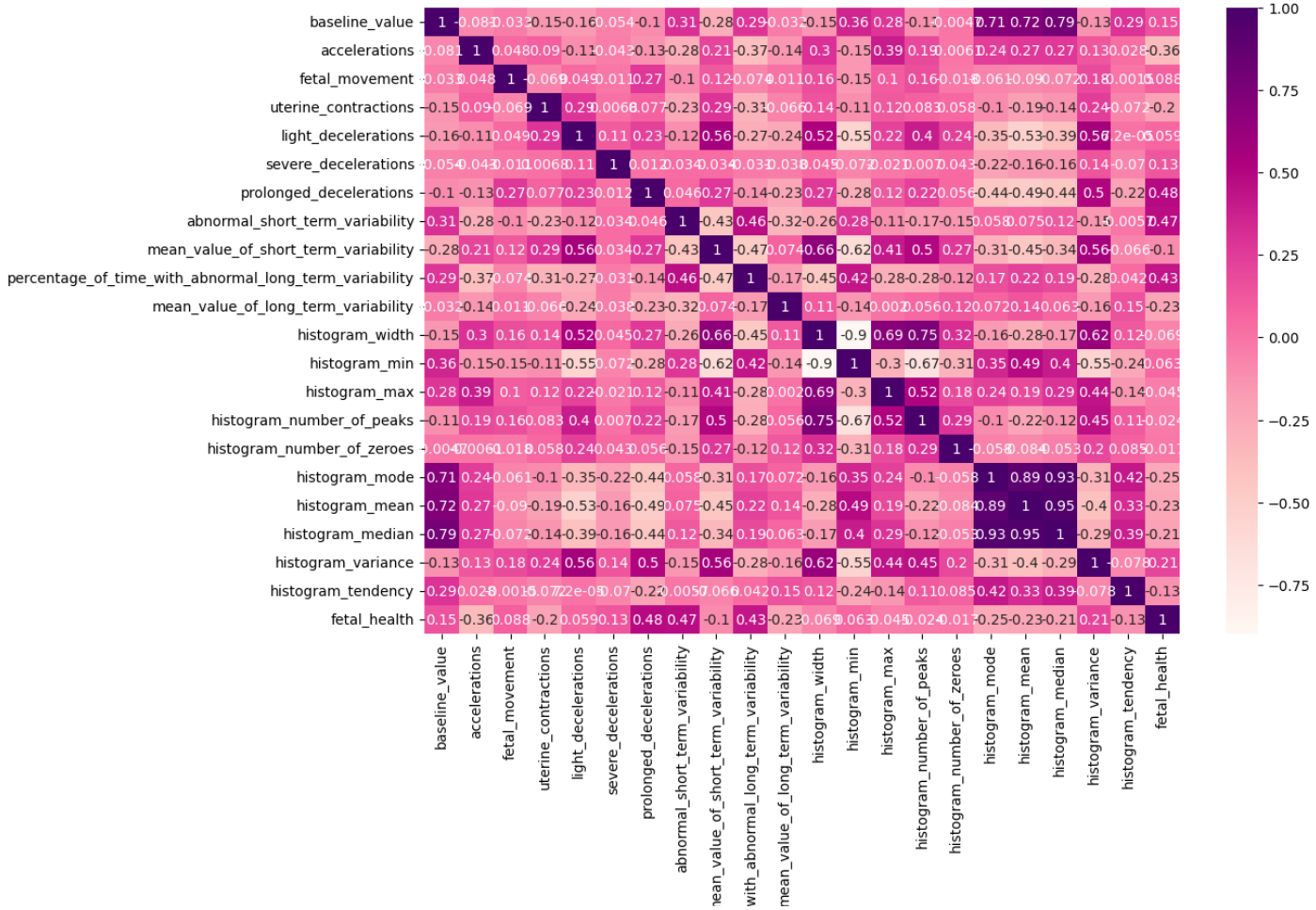
The data didn't have any null or missing values and was fairly clean for us to start exploring variables and generate insights



The count plot of the target variable (fetal_health) reveals an imbalance in the data, a situation that can potentially lead to inaccurate classification.

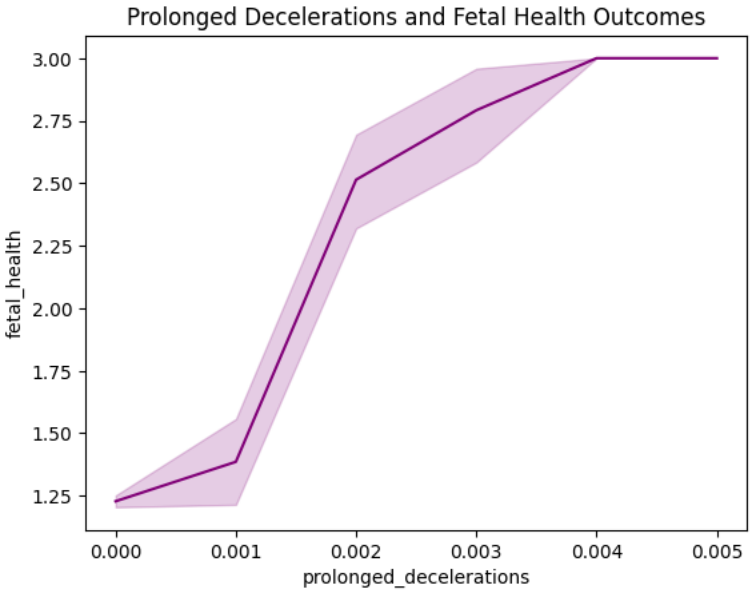


All the attributes are midly skewed and are normally distributed except the features "light_decelerations", "precentage_of_time_with_abnormal_long_term_variability" etc.

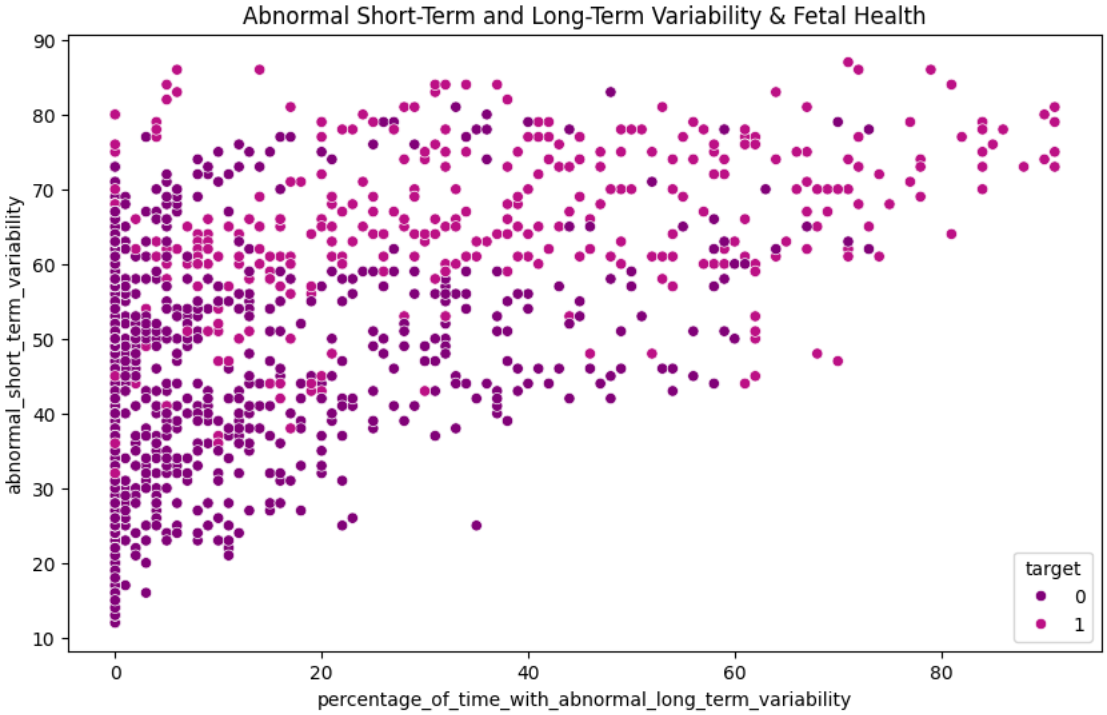


Examining the correlation matrix reveals that "accelerations," "prolongued_decelerations" "abnormal_short_term_variability","percentage_of_time_with_abnormal_long_term_variability" and "mean_value_of_long_term_variability" are highly correlated with fetal_health. To gain deeper insights into these features concerning fetal_movement on the y-axis, we aim to obtain a more detailed understanding of trends indicative of fetal health.

Data Visualizations

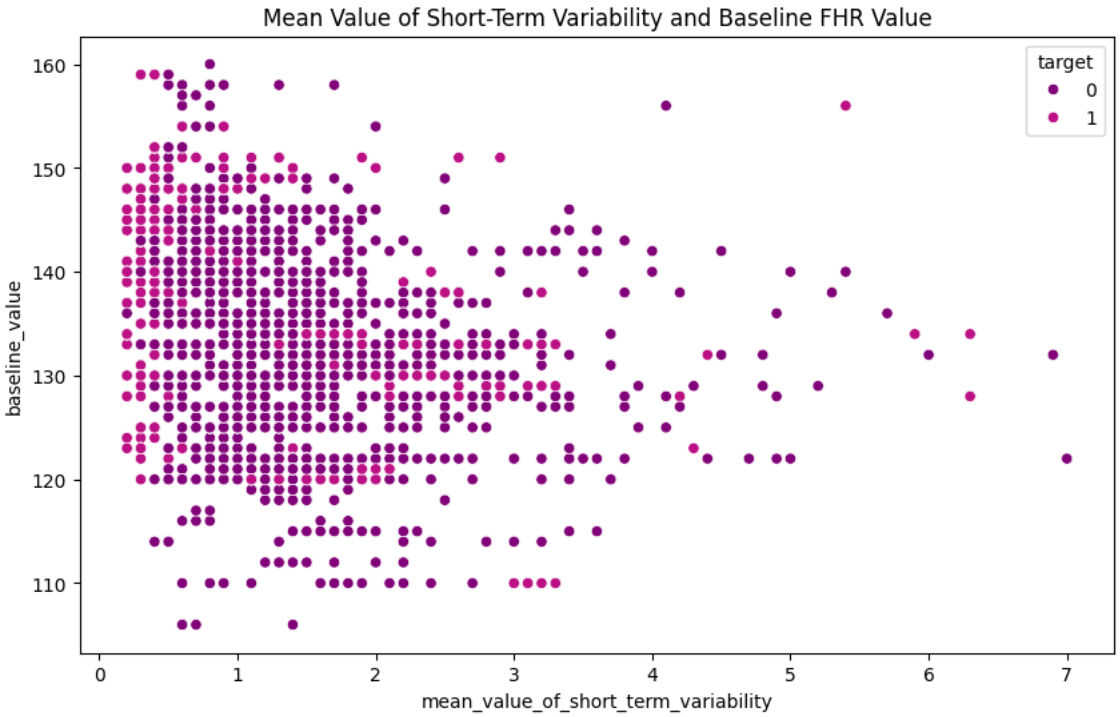


This lineplot clearly shows the relationship between prolonged decelerations of fetal heart rate and fetal health outcome. The longer amount of time observed with prolonged decelerations, the more at-risk the health outcome was likely to be.

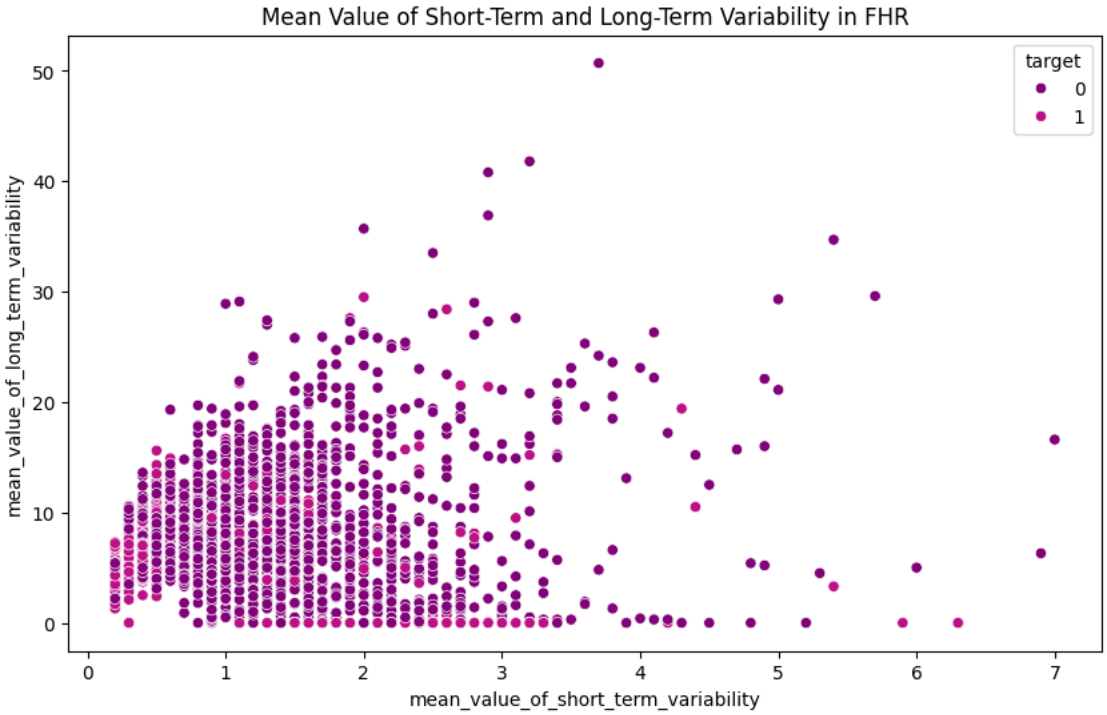


The light red points represent the At Risk class and the dark red points represent the Normal class. From this scatter plot, it can be seen that the points with higher values for percentage of time

with short- and long- term variability are predominantly of the At Risk class, with only a few Normals peppered in there.

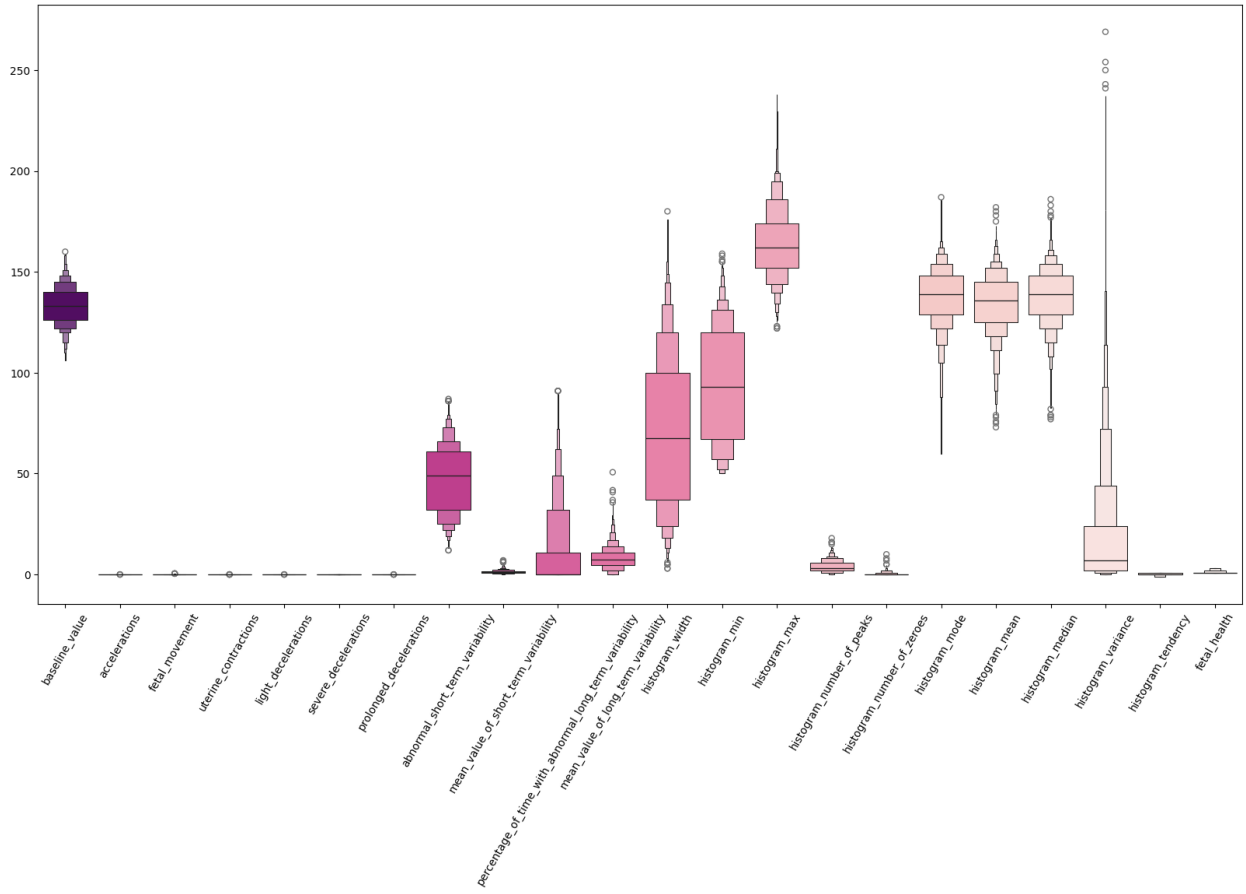


When comparing baseline FHR value and mean value of short term variability, it is not as easy to distinguish a clear relationship between the classes.

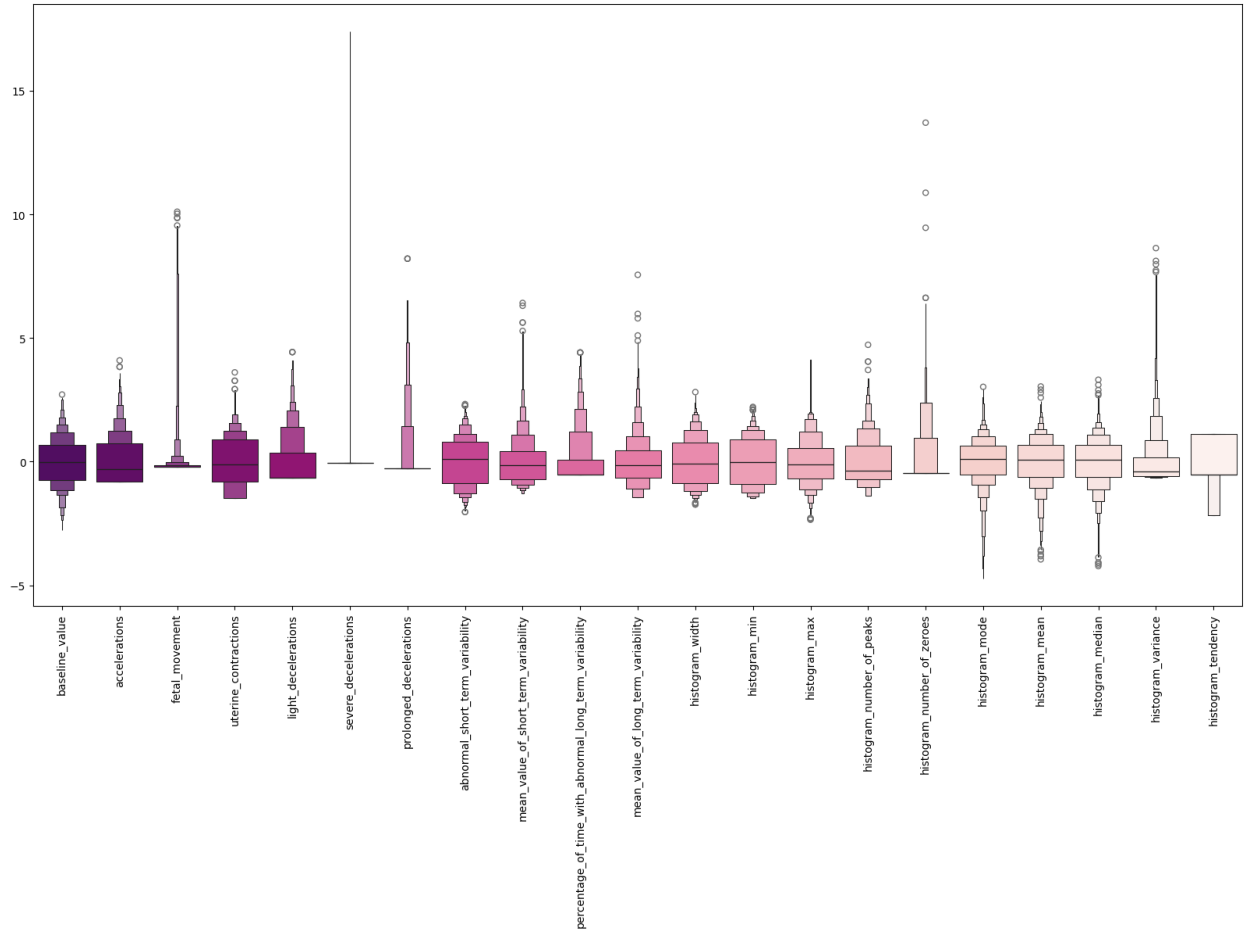


It is again difficult to distinguish the pattern here. It is notable that there is no clear way to discern the two classes based on these features alone.

Data Processing



The above plot shows the range of our feature attributes. All the features are in different ranges. To fit this in a model we must scale it to the same range.

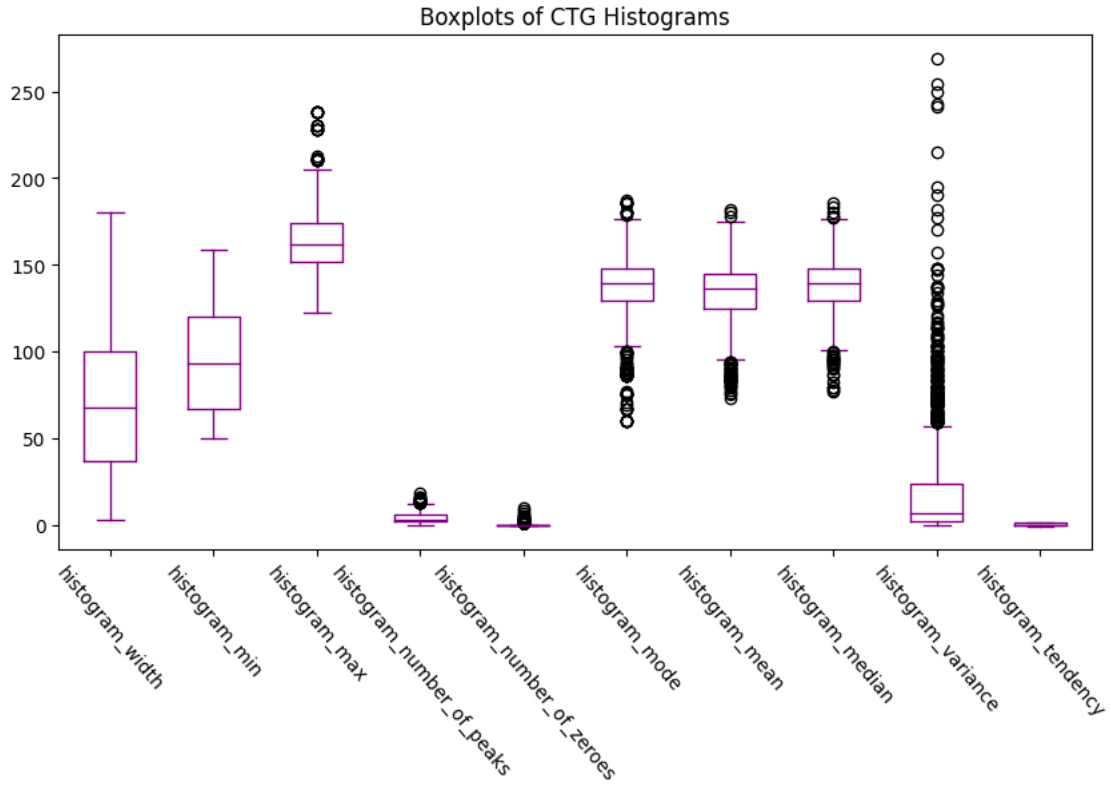


We use the MinMaxScaler to scale our data to the same range.

	count	mean	std	min	25%	50%	75%	max
baseline_value	2126.0	1.069490e-15	1.000235	-2.775197	-0.742373	-0.030884	0.680604	2.713428
accelerations	2126.0	-4.010589e-17	1.000235	-0.822388	-0.822388	-0.304881	0.730133	4.093929
fetal_movement	2126.0	-1.336863e-17	1.000235	-0.203210	-0.203210	-0.203210	-0.138908	10.106540
uterine_contractions	2126.0	-1.336863e-16	1.000235	-1.482465	-0.803434	-0.124404	0.894142	3.610264
light_decelerations	2126.0	-5.347452e-17	1.000235	-0.638438	-0.638438	-0.638438	0.375243	4.429965
severe_decelerations	2126.0	6.684315e-18	1.000235	-0.057476	-0.057476	-0.057476	-0.057476	17.398686
prolonged_decelerations	2126.0	1.336863e-17	1.000235	-0.268754	-0.268754	-0.268754	-0.268754	8.208570
abnormal_short_term_variability	2126.0	-7.352747e-17	1.000235	-2.035639	-0.872088	0.116930	0.815060	2.327675
mean_value_of_short_term_variability	2126.0	6.684315e-17	1.000235	-1.282833	-0.716603	-0.150373	0.415857	6.417893
percentage_of_time_with_abnormal_long_term_variability	2126.0	-5.347452e-17	1.000235	-0.535361	-0.535361	-0.535361	0.062707	4.412293
mean_value_of_long_term_variability	2126.0	2.406354e-16	1.000235	-1.455081	-0.637583	-0.139975	0.464263	7.555172
histogram_width	2126.0	-3.007942e-17	1.000235	-1.731757	-0.858765	-0.075640	0.758838	2.812936
histogram_min	2126.0	-4.679021e-17	1.000235	-1.474609	-0.899376	-0.019608	0.893996	2.213648
histogram_max	2126.0	-1.203177e-16	1.000235	-2.342558	-0.670314	-0.112899	0.555999	4.123453
histogram_number_of_peaks	2126.0	-1.671079e-16	1.000235	-1.379664	-0.701397	-0.362263	0.655137	4.724738
histogram_number_of_zeroes	2126.0	2.757280e-17	1.000235	-0.458444	-0.458444	-0.458444	-0.458444	13.708003
histogram_mode	2126.0	1.069490e-16	1.000235	-4.729191	-0.516077	0.094519	0.644055	3.025381
histogram_mean	2126.0	-6.684315e-16	1.000235	-3.951945	-0.616458	0.089126	0.666422	3.039749
histogram_median	2126.0	2.673726e-16	1.000235	-4.223849	-0.628514	0.062897	0.685166	3.312527

After scaling, the statistics of the data is more uniform.

DIMENSION REDUCTION AND VARIABLE SELECTION



Many features in the dataset are related to the histogram generated during Cardiotocography (CTG). However, the interpretation of these histogram measurements appears to lack clarity and intuition. Before deciding to remove these columns, we sought to understand how these measurements might influence the prediction of fetal health outcomes.

MODEL EXPLORATION AND MODEL SELECTION

Here's an overview of classification models (Logistic Regression, Decision Tree, Random Forest, Support Vector Machines, and KNN) for our fetal health classification project:

1. Logistic Regression:

- **Overview**: Logistic Regression is a statistical method for binary classification that models the probability of the outcome using the logistic function. It's widely used for its simplicity and interpretability, making it a popular choice in various fields.

- **Pros**:
  - Simple and easy to understand.
  - Provides probabilities for class membership.
  - Works well when the relationship between features and the log-odds of the outcome is approximately linear.

- **Cons**:

- Assumes a linear relationship, which may not hold in all cases.
- Limited to binary classification (extensions exist for multiclass).
- Sensitive to outliers.

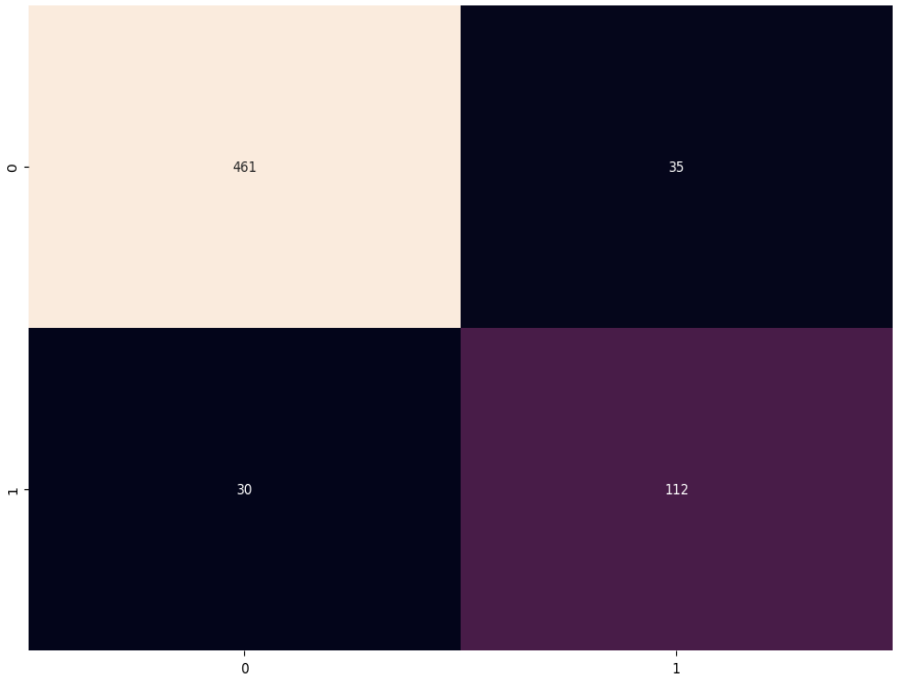
Expected Results: Logistic Regression is suitable for scenarios where the relationship between features and the outcome is relatively linear and when interpretability is crucial. It's a good starting point for binary classification tasks, especially when there's a need for probability estimates. However, its performance may be limited in cases with complex, nonlinear relationships.

Actual Results:

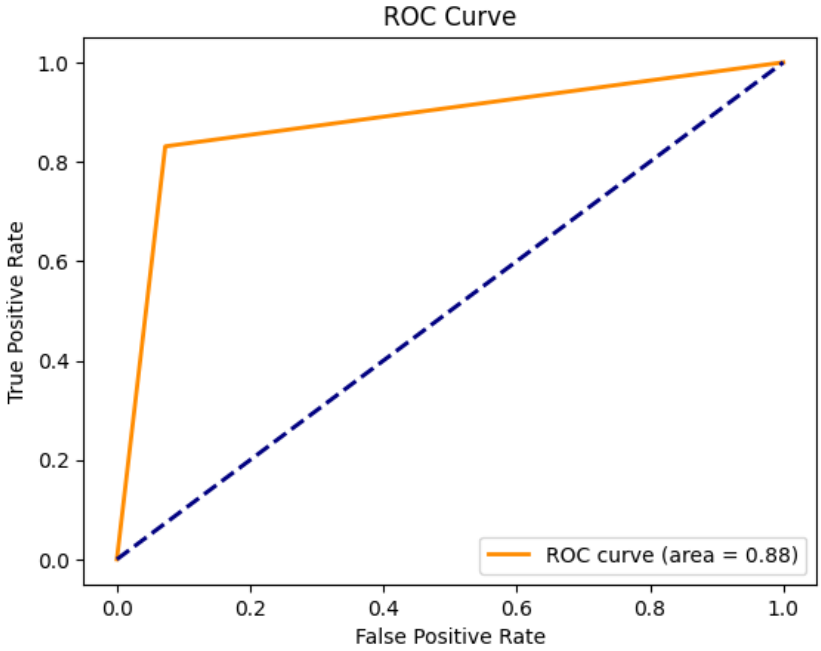
Classification Report:

	precision	recall	f1-score	support
0	0.95	0.93	0.94	496
1	0.77	0.83	0.80	142
accuracy			0.91	638
macro avg	0.86	0.88	0.87	638
weighted avg	0.91	0.91	0.91	638

Confusion Matrix:



ROC/AUC Curve:



2. **Decision Tree:**

- **Overview**: Decision Trees recursively split the data based on features to make decisions. They are intuitive, easy to interpret, and can handle both numerical and categorical data.
- **Pros**:
  - Intuitive and easy to understand.
  - Minimal data preprocessing required.
  - Can handle mixed data types.

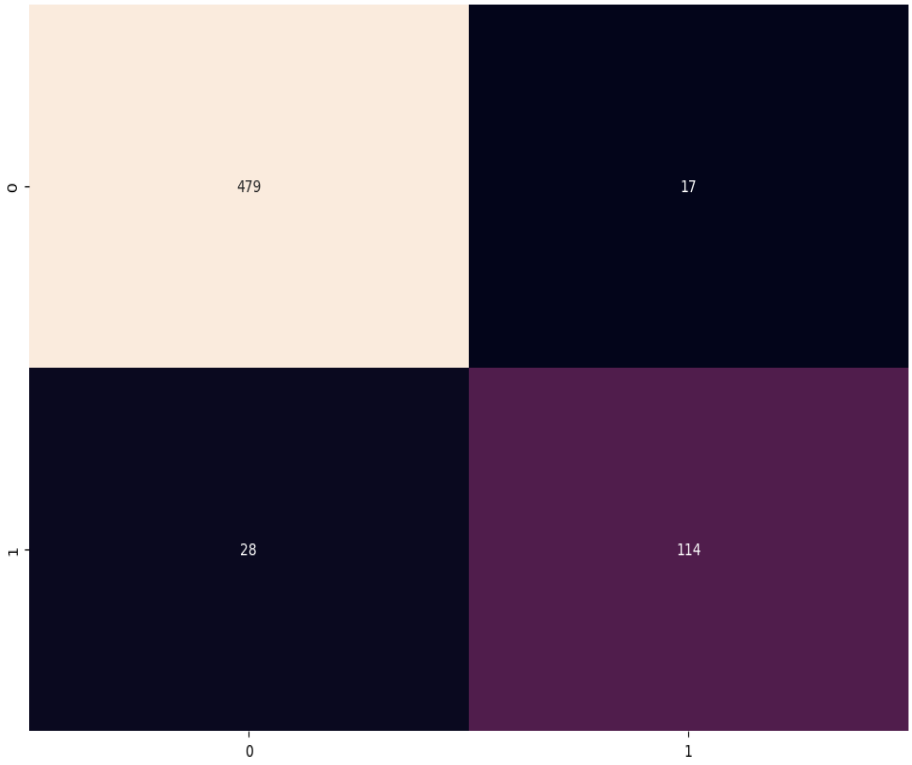
- **Cons**:
  - Prone to overfitting, especially with deep trees.
  - Sensitive to small variations in the data.
  - May struggle with capturing relationships between distant features.

Expected Results: Decision Trees are simple and suitable for quick insights into the data structure. Regularization and pruning are essential to prevent overfitting.

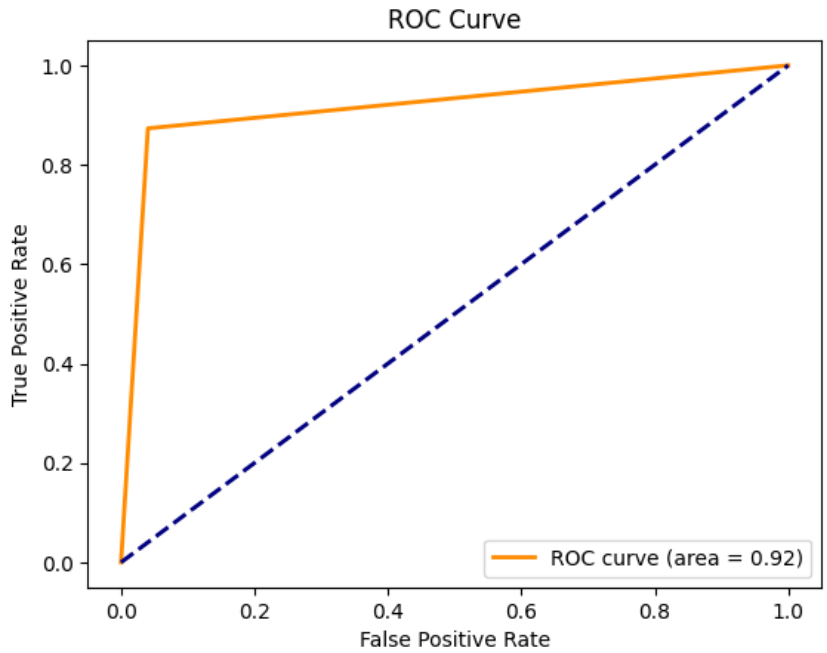
Actual Results:

```
Accuracy: 0.9404388714733543
Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.96      0.96       496
           1       0.86      0.87      0.87       142

    accuracy                           0.94       638
   macro avg       0.91      0.92      0.91       638
weighted avg       0.94      0.94      0.94       638
```

Confusion Matrix:



ROC/AUC Curve:



3. **Random Forest:**

- **Overview**: Random Forest is an ensemble learning method that builds multiple Decision Trees and combines their predictions. It aims to improve predictive accuracy and reduce overfitting.

- **Pros**:
  - High predictive accuracy through ensemble learning.
  - Reduces overfitting compared to a single Decision Tree.
  - Provides feature importance for interpretability.

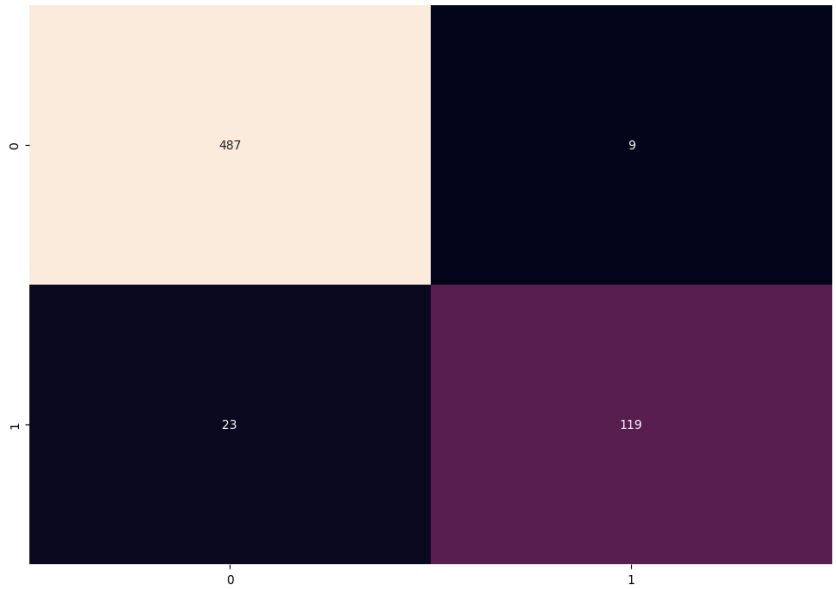
- **Cons**:
  - Less interpretable compared to individual Decision Trees.
  - Can be computationally expensive, especially for large datasets.
  - Requires careful tuning of hyperparameters.

Expected Results: Random Forests strike a balance between simplicity and accuracy. They are effective in capturing complex relationships and are suitable when interpretability is important, but a more advanced model is needed compared to a standalone Decision Tree.

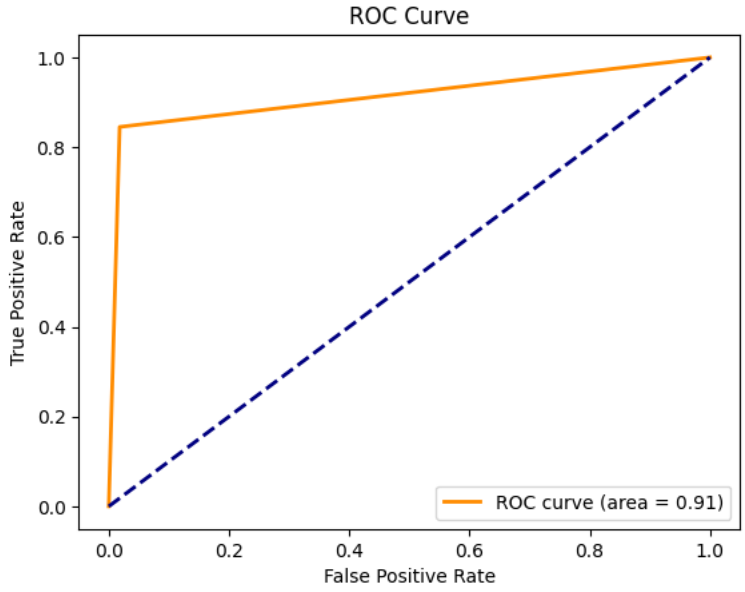
Actual Results:

```
Accuracy: 0.95141065830721
Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.98      0.97       496
           1       0.93      0.85      0.89       142

    accuracy                           0.95       638
   macro avg       0.94      0.91      0.93       638
weighted avg       0.95      0.95      0.95       638
```

Confusion Matrix:



ROC/AUC Curve:



4. **Support Vector Machines (SVM):**

- **Overview**: SVM is a powerful classification algorithm that finds the optimal hyperplane to separate data into classes. It is effective in high-dimensional spaces.
- **Pros**:
  - Effective in high-dimensional spaces.
  - Versatile, with different kernel functions available.
  - Robust to overfitting, especially in high-dimensional spaces.

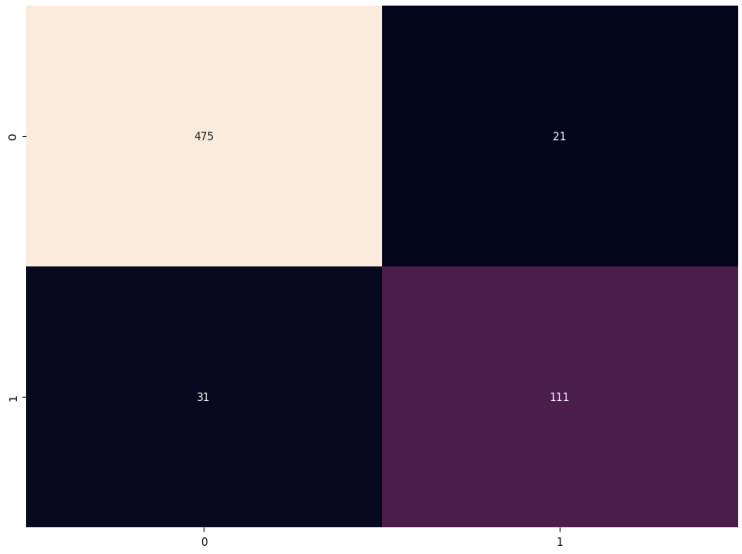
- **Cons**:
  - Memory-intensive, especially for large datasets.
  - Sensitive to the choice of the kernel and parameters.
  - Can be computationally expensive during training.

Expected Results: SVMs are powerful for complex relationships and can handle high-dimensional data well. They perform better when there is a clear margin of separation.

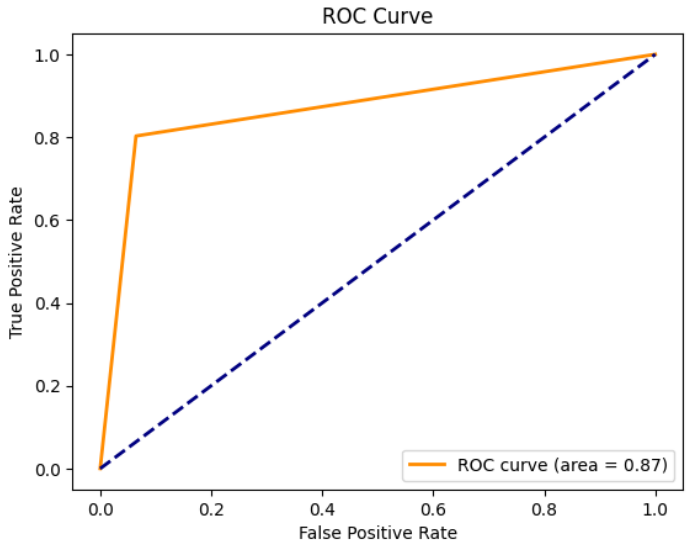
Actual Results:

```
Accuracy: 0.9059561128526645
Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.94      0.94       496
           1       0.78      0.80      0.79       142

    accuracy                           0.91       638
   macro avg       0.86      0.87      0.87       638
weighted avg       0.91      0.91      0.91       638
```

Confusion Matrix:



ROC/AUC Curve:



5. **k-Nearest Neighbors (kNN):**

- Overview: k-Nearest Neighbors is a non-parametric and instance-based learning algorithm. It classifies new data points based on the majority class of their k nearest neighbors in the feature space. The choice of k and the distance metric are crucial parameters in kNN.

- Pros:

- Simple and easy to implement.
- No assumptions about the underlying data distribution.

- Effective for both classification and regression tasks.
- Adapts well to local patterns and can handle non-linear relationships.

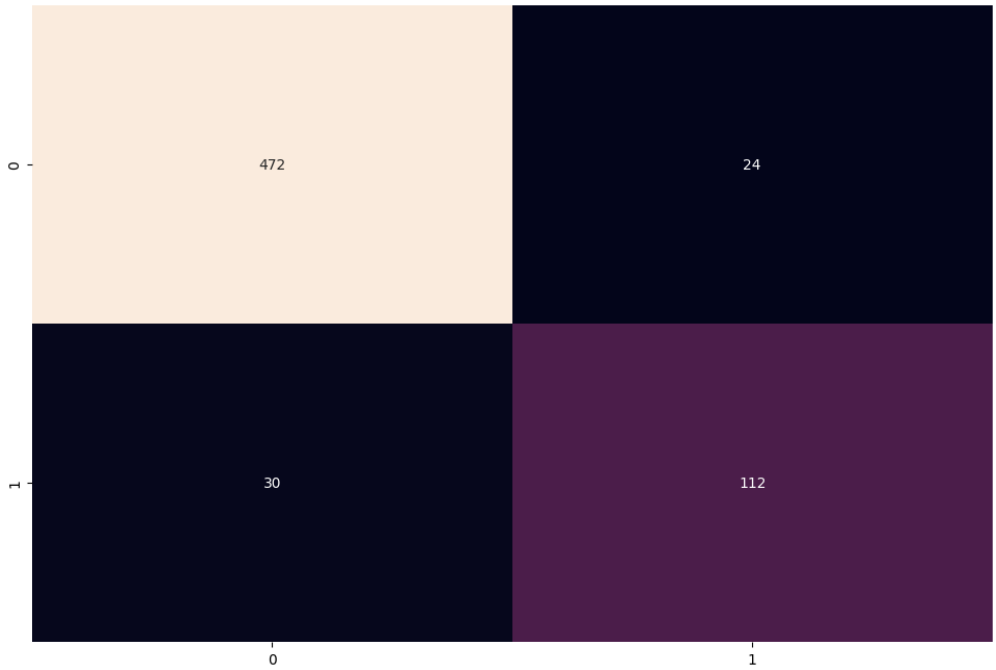
- Cons:
- Computationally expensive, especially with large datasets.
- Sensitive to irrelevant or redundant features.
- Performance can degrade if the dataset has a high dimensionality.
- The choice of an appropriate distance metric is crucial.

Expected Results: kNN is a versatile algorithm that can be effective in scenarios where local patterns are important, and the decision boundary is not globally smooth. It's suitable for situations where the dataset is not too large, and computational efficiency is not a primary concern. Careful consideration of the distance metric and tuning the value of k is essential for optimal performance.

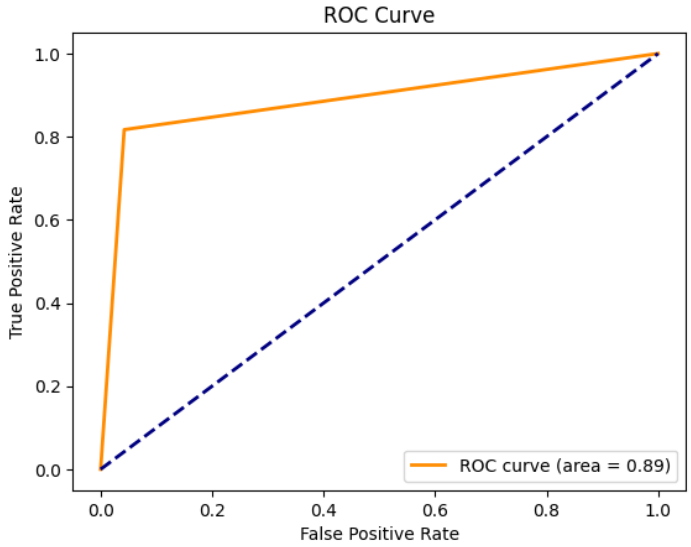
Actual Results:

```
Accuracy: 0.9263322884012539
Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.96      0.95       496
           1       0.85      0.82      0.83       142

    accuracy                           0.93       638
   macro avg       0.90      0.89      0.89       638
weighted avg       0.93      0.93      0.93       638
```

Confusion Matrix:



ROC/AUC Curve:



OVERALL SUMMARY

	Accuracy	Precision	Recall	F1 Score	AUC
Logistic Regression	0.905956	0.909420	0.905956	0.907287	0.959308
K-Nearest Neighbors	0.926332	0.925538	0.926332	0.925857	0.939871
Random Forest	0.951411	0.950870	0.951411	0.950559	0.988173
Support Vector Machine	0.905956	0.906973	0.905956	0.906419	0.957988
Decision Tree	0.940439	0.940760	0.940439	0.940587	0.916458

Model Selection

Based on the results, we have decided to select random forest as our model of choice. This model has given us with the best recall scores and the least number of mis-classifications w.r.t the number of False Negative cases.

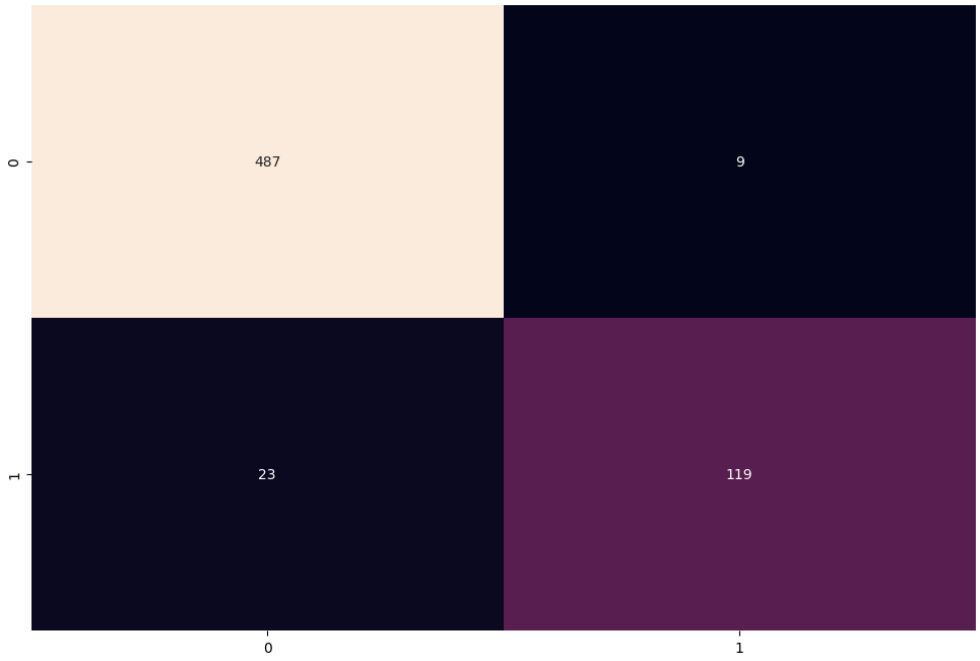
MODEL PERFORMANCE EVALUATION – RANDOM FOREST

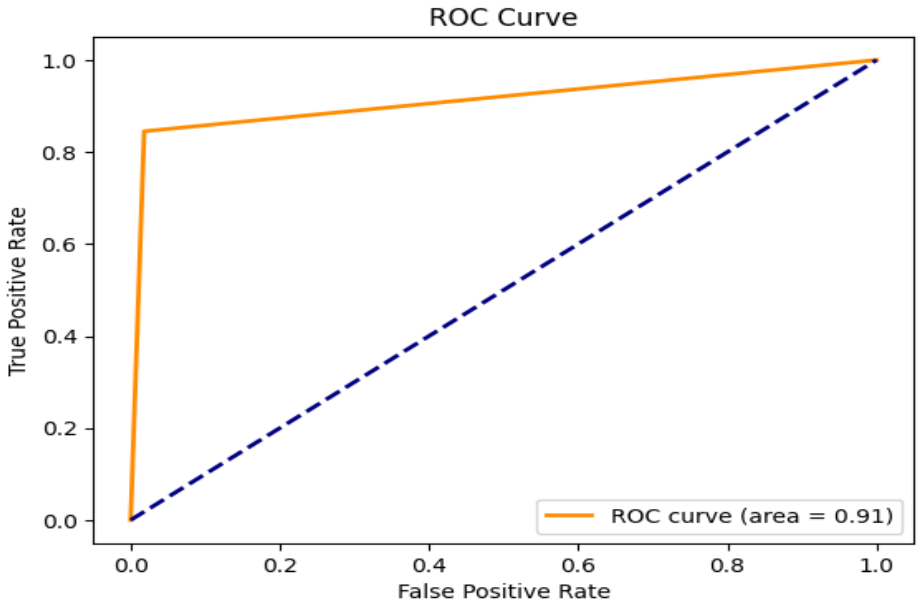
```
Accuracy: 0.95141065830721
Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.98      0.97       496
           1       0.93      0.85      0.89       142

    accuracy                           0.95       638
   macro avg       0.94      0.91      0.93       638
weighted avg       0.95      0.95      0.95       638
```

The model has performed well achieving an accuracy of 95%

PERFORMANCE VISUALIZATIONS

The following confusion matrix and AUC/ROC curves visualizes the performance of the model, indicating strong results while comparing the predicted values to the actual values.





REFERENCES

Dataset was obtained from the UC Irvine's ML repository - https://archive.ics.uci.edu/dataset/193/cardiotocography

The original study for automated analysis of CTG data can be found at - https://onlinelibrary.wiley.com/doi/10.1002/1520-6661(200009/10)9:5%3C311::AID-MFM12%3E3.0.CO;2-9.