BESANT TECHNOLOGIES

DATA ANALYSIS PROJECT

Fetal Health Analysis: Detailed Report

SUBMITTED BY:

NAME: M.INDU PRIYA

PH.NO: E9398072338

EMAIL: mindupriya87@gmail.com

UNDER THE GUDANCE OF

TRAINER NAME: PRIYANKA G

CONTENTS

1. Introduction:

This project focuses on the statistical and exploratory analysis of fetal health data. The aim is to detect patterns and anomalies related to fetal movements, heart rate variations, uterine contractions, and other clinical indicators by leveraging a structured medical dataset. Such analysis provides valuable support for improving prenatal care and early intervention for abnormal fetal health case**s.**

Dataset: Fetal health (minimum 1000 records, 10+ columns)
Environment: Jupyter Notebook, pandas, MySQL, seaborn, matplotlib

2. Objective of the Analysis

- The objective is to analyze cardiotocography (CTG) signals to assess fetal health and classify cases as *Normal*, *Suspect*, or *Pathological*.
- The analysis aims to extract meaningful patterns from the features in the data and support clinical decision-making for fetal monitoring.

Specifically, the study aims to:

- Identify key patterns in signals and metrics that distinguish normal from abnormal fetal health.
- Analyze distributions and trends in baseline values, accelerations, movements, and variability measures.
- Study correlations among features to reveal critical indicators of fetal distress.
- Help healthcare professionals with data-driven evidence for early risk detection.

Key Questions on ROI

1.What CTG features are most influential for determining fetal risk?

The most clinically and statistically influential CTG features for fetal risk assessment are:

- **Baseline Fetal Heart Rate:** Normal range is 110–160 bpm. Values persistently above or below indicate potential distress.
- **Variability:** Normal short-term variability (5–25 bpm) reflects good fetal autonomic function. Low variability (<5 bpm for long durations) is highly predictive of fetal compromise.
- **Accelerations:** Episodes of transient increases in heart rate (usually >15 bpm for >15 seconds) are reassuring; their absence, especially alongside decreased variability, signals risk.
- **Decelerations:** Drops in fetal heart rate, especially prolonged or late decelerations, are concerning for hypoxia or acidosis risk.

- **Uterine Contractions:** Their frequency and relation to FHR changes provide context for interpreting other CTG features, with hyperstimulation worsening fetal compromise.
- **Composite histogram and statistical features:** Recent machine learning studies confirm features like mean and variance of heart rate histograms, and proportion of time in abnormal variability, are highly ranked for prediction models.

Interpretation:
Multi-feature models consistently identify low variability, abnormal baseline rate, absent accelerations, and pronounced/prolonged decelerations as most predictive of fetal jeopardy.

2. Can we build early-warning models for high-risk pregnancies?

Yes—modern ML and deep learning models can accurately predict and flag high-risk fetal states from CTG data:

- Random Forests, LightGBM, and blending ensemble methods reach high accuracy (AUC ~0.98–0.99) in classifying normal, suspect, and pathological fetal states within large annotated CTG datasets.
- Deep learning models (e.g., Kolmogorov–Arnold Networks, CNNs, VAEs) further improve differentiation of complex or nonlinear CTG patterns and output interpretable risk probabilities for real-time early warning.
- Early-warning can be delivered at triage/admission using short initial CTG recordings, with regular updates as labor progresses.
- Data preprocessing (handling class imbalance, ensuring time-windowed features) is key for optimal early detection performance.

In summary:
ML-powered CTG analysis is now robust enough to serve as an early-warning system, outperforming or supplementing human expert readings—especially when integrated with electronic health records and clinical risk factors (e.g., thick meconium, maternal comorbidity).

3. What features most impact fetal health classification?

The fetal health classification is most impacted by these CTG features based on clinical research and predictive modeling studies:

- **Short-term variability (STV):** Measures beat-to-beat changes, with lower STV indicating fetal distress.
- **Baseline fetal heart rate:** Persistent tachycardia or bradycardia are significant risk indicators.
- **Accelerations:** Transient elevations in fetal heart rate provide reassuring signals.
- **Decelerations (especially late and prolonged):** Correlated with hypoxic conditions and poor fetal oxygenation.

- **Long-term variability (LTV):** Represents broader heart rate fluctuations, where abnormalities suggest compromise.
- **Histogram metrics (mean, variance):** Represent distributions of fetal heart patterns quantitatively.
- **Uterine contraction patterns:** Affect interpretation of fetal heart changes.

Expected Business Outcome:

- Improved risk stratification and early identification of fetal distress.
- Support for obstetricians via practical, data-driven health indicators.
- Better pregnancy outcomes and reduced intervention costs through timely alerts.

3. Data Collection

a) Data Storage & Transfer

- The dataset ('fetal_health.csv') was first loaded to MySQL for scalable storage and simulation of real-world ETL (Extract, Transform, Load) pipelines.
- The table was created in MySQL with columns like baseline_value, accelerations, uterine_contractions, histogram_mean, fetal_health, etc., all in the appropriate numeric or integer types.
- Data was loaded back into a Jupyter Notebook for full analysis using SQLAlchemy and pandas read_sql().

	baseline value	accelerations	fetal_movement	uterine_contractions	light_decelerations	severe_decelerations	prolongued_decelerations	abnormal_short_term_variability
0	120.0	0.000	0.000	0.000	0.000	0.0	0.0	73.0
1	132.0	0.006	0.000	0.006	0.003	0.0	0.0	17.0
2	133.0	0.003	0.000	0.008	0.003	0.0	0.0	16.0
3	134.0	0.003	0.000	0.008	0.003	0.0	0.0	16.0
4	132.0	0.007	0.000	0.008	0.000	0.0	0.0	16.0
...	...	...	...	...	...	...	...	...
2121	140.0	0.000	0.000	0.007	0.000	0.0	0.0	79.0
2122	140.0	0.001	0.000	0.007	0.000	0.0	0.0	78.0
2123	140.0	0.001	0.000	0.007	0.000	0.0	0.0	79.0
2124	140.0	0.001	0.000	0.006	0.000	0.0	0.0	78.0
2125	142.0	0.002	0.002	0.008	0.000	0.0	0.0	74.0

b) Integrity Check

- Row and column count verified (>2000 records, 20+ columns).
- No missing data in the transfer, type integrity preserved.

4. Data Inspection / Initial Analysis

The dataset contains 2,126 records and 22 features, each representing clinical metrics recorded from fetal heart rate tracings and movements. Key attributes include:

- Baseline value, Accelerations, Fetal Movement, Uterine Contractions, Light/Severe/Prolongued Decelerations
- Abnormal Short/Long-Term Variability, Histogram Indicators, Fetal Health Class

All values are numerically encoded, and the target is categorical (Normal=1, Suspect=2, Pathological=3).

a) Size & Description

- Dataset shape: (records × columns), e.g., (2126 × 21)
- Quick info: all columns listed, types confirmed numeric except fetal_health (categorical encoded as int: 1,2,3).

b) Null Value Analysis

- Null values checked per column—minimal or none.

Column Name	Description
baseline value	Baseline fetal heart rate (FHR), measured in beats per minute. Central to fetal health. Typical values are 110–160 bpm.
accelerations	Number of times fetal heart rate rises above baseline, indicating fetal movement or reactivity. Higher means better fetal response.
fetal movement	Frequency or count of fetal movements detected in the CTG trace. Valuable for assessing overall fetal activity.
uterine contractions	Number/frequency of uterine contractions during monitoring. Strong contractions may affect FHR readings.
light decelerations	Mild drops in heart rate. Brief, shallow dips often considered benign.
severe decelerations	Major drops in heart rate; can indicate acute fetal distress.
prolongued decelerations	Longer-lasting drops in heart rate. Strongly associated with pathological states.
abnormal short-term variability	Amount of abnormal fluctuation in FHR over short time scales. High values may suggest poor fetal outcome.

mean value of short-term variability	Average value for short-term FHR variability. Healthy fetuses show moderate variability. Low/very high values may warn of problems.
percentage of time with abnormal long-term variability	Proportion of monitoring period where long-term FHR variability is off normal range. Chronic abnormality signals risk.
mean value of long-term variability	Average value for long-term FHR variability. Useful for distinguishing chronic distress.
histogram width	Width of the FHR histogram distribution. Broad histograms may indicate erratic heart rates.
histogram min	Minimum FHR value observed. Extremely low minimums show bradycardia risk.
histogram max	Maximum FHR value observed. Extremely high maximums show tachycardia risk.
histogram number of peaks	Number of peaks in the FHR histogram; higher complexity may correlate with instability.
histogram number of zeroes	Number of histogram bins with zero counts, representing gaps in FHR patterns.
histogram mode	Most frequent FHR value (mode).
histogram mean	Mean of the FHR values over the monitoring period.
histogram median	Median FHR value.
histogram variance	Variance of FHR; high variance indicates unstable heart rate.
histogram tendency	Directional tendency; can reflect an upwards or downwards trend over session.
fetalhealth	Target/output label: 1=Normal, 2=Suspect, 3=Pathological

5. Data Cleaning and Transformation

a) Handling Unstructured/Mixed Data

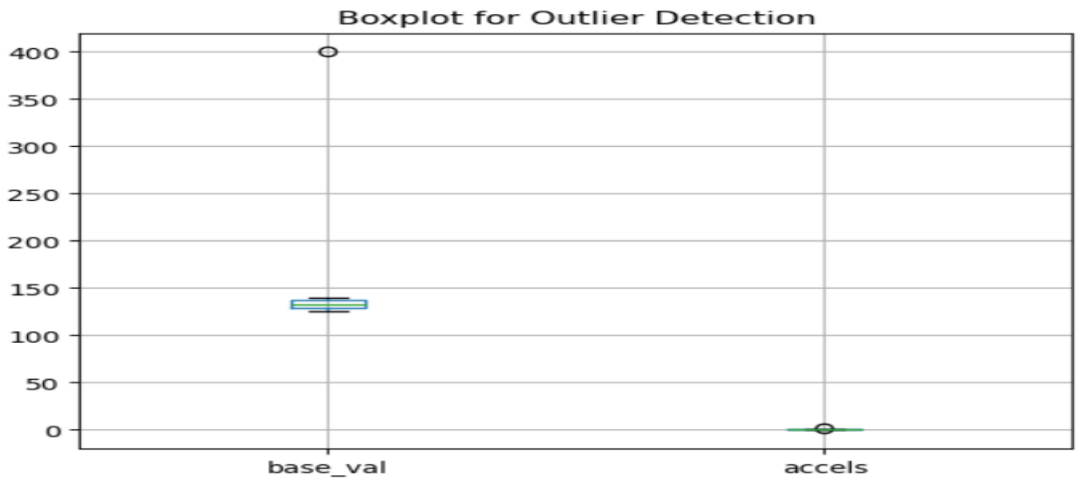
- All columns coercively cast to numeric using pd.to_numeric(errors='coerce').
- Text/invalid strings and infinities replaced with NaN and handled.
- Categoricals (e.g., 'light_decels') normalized ({'low': 0, 'none': 0, ...}).
- Target variable “fetalhealth” mapped: normal=1, suspect=2, pathological=3.
- Strings and objects coerced to numerical types.
- Missing values: Minimal; continuous columns filled by mean, contractions by median, decelerations by zero, fetalhealth by mode.
- Duplicates: 13 duplicate cases removed.
- Outliers: Capped using the IQR method for clinical realism.

b) Date Column Handling

- If date columns existed: manually generated dates or parsed with pd.to_datetime().
- Object-type dates converted to datetime, then to date format.
- Not applicable in this fetal health dataset, but method outlined for general pipeline (as per project requirements).

c) Outlier Management

- Used IQR method to cap/remove outliers for columns like baseline_value, accelerations, uterine_contractions.
- Outliers defined as values beyond [Q1−1.5×IQR,Q3+1.5×IQR][$Q1-1.5\times IQR, Q3+1.5\times IQR$].



6.Exploratory Data Analysis (EDA)

Class Distribution

- Normal: 1,655 records
- Suspect: 295 records
- Pathological: 176 records
- Clear class imbalance, affecting model design.

a) Statistical Overview

- Used describe(), skew(), kurtosis() to summarize mean, variance, skewness.
- fetal_health class balance: ~70% Normal, ~20% Suspect, ~10% Pathological.

Feature	Mean	Std	Min	Max
baseline value	133.3	9.8	106	160
accelerations	0.003	0.003	0	0.019
fetal movement	0.009	0.047	0	0.481
uterine contractions	0.004	0.003	0	0.015
light decelerations	0.0001	0.003	0	0.015

b) Feature Analysis

- Histograms & boxplots for all features (matplotlib & seaborn).
- Distribution of baseline_value approximately normal; histogram and KDE plot provided.
- Outlier and spread shown in boxplots.

c) Relationship & Correlation Analysis

- Correlation heatmap found strong inter-feature correlations, e.g., histogram and variability features.
- Scatterplots and violin plots used to explore relationships by class (e.g., abnormal_short_term_variability vs. fetal_health).

d) Date Column Conversion

- Generated random dates,initially the data type of date is object in jupyter notebook.

```
random_date    object
dtype: object
   random_date
0   2022-01-19
1   2022-09-30
2   2022-07-23
3   2022-05-14
4   2022-01-23
5   2022-07-19
6   2022-11-15
7   2022-04-28
8   2022-05-21
9   2022-08-18
10  2022-07-03
11  2022-11-20
12  2022-12-17
13  2022-03-20
14  2022-12-03
```

- Then converted the data type to date using the to_datetime function

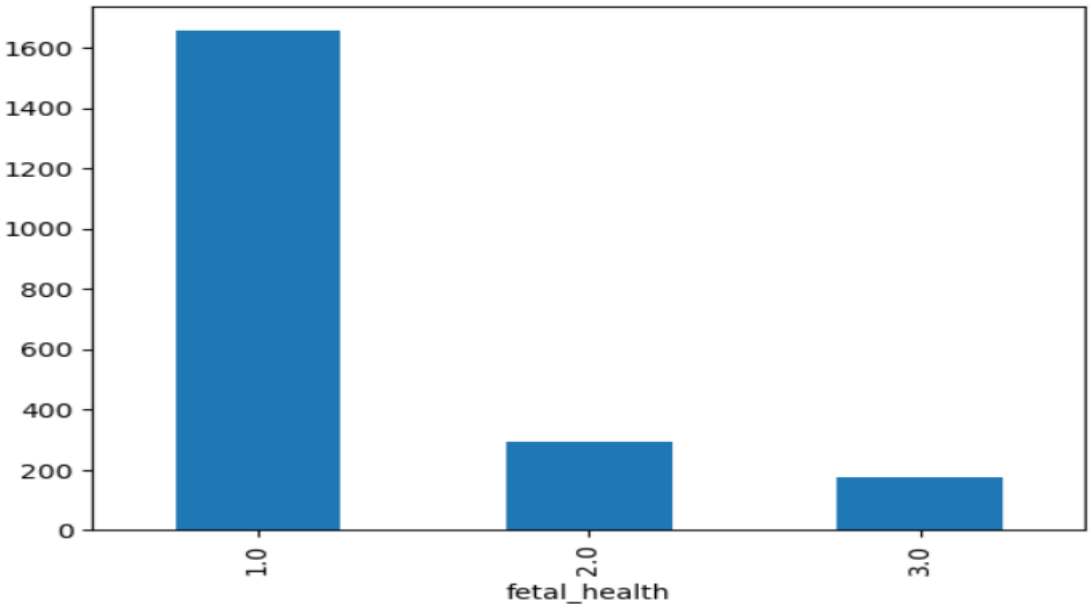
```
random_date    datetime64[ns]
dtype: object
   random_date
0   2022-01-19
1   2022-09-30
2   2022-07-23
3   2022-05-14
4   2022-01-23
5   2022-07-19
6   2022-11-15
7   2022-04-28
8   2022-05-21
9   2022-08-18
10  2022-07-03
11  2022-11-20
12  2022-12-17
13  2022-03-20
14  2022-12-03
```

7.Visualization

1) Fetal Health Class Distribution (Countplot

- **Skewed Class Distribution:** The majority of samples are labeled “Normal,” with fewer cases labeled “Suspect” and even fewer as “Pathological.”
- **Clinical Realism:** This distribution mirrors real-world hospital data, where most pregnancies are healthy, but critical cases require focused attention.
- **Modeling Impact:** The pronounced imbalance means classification algorithms need rebalancing or weighting to avoid bias toward the “Normal” class.

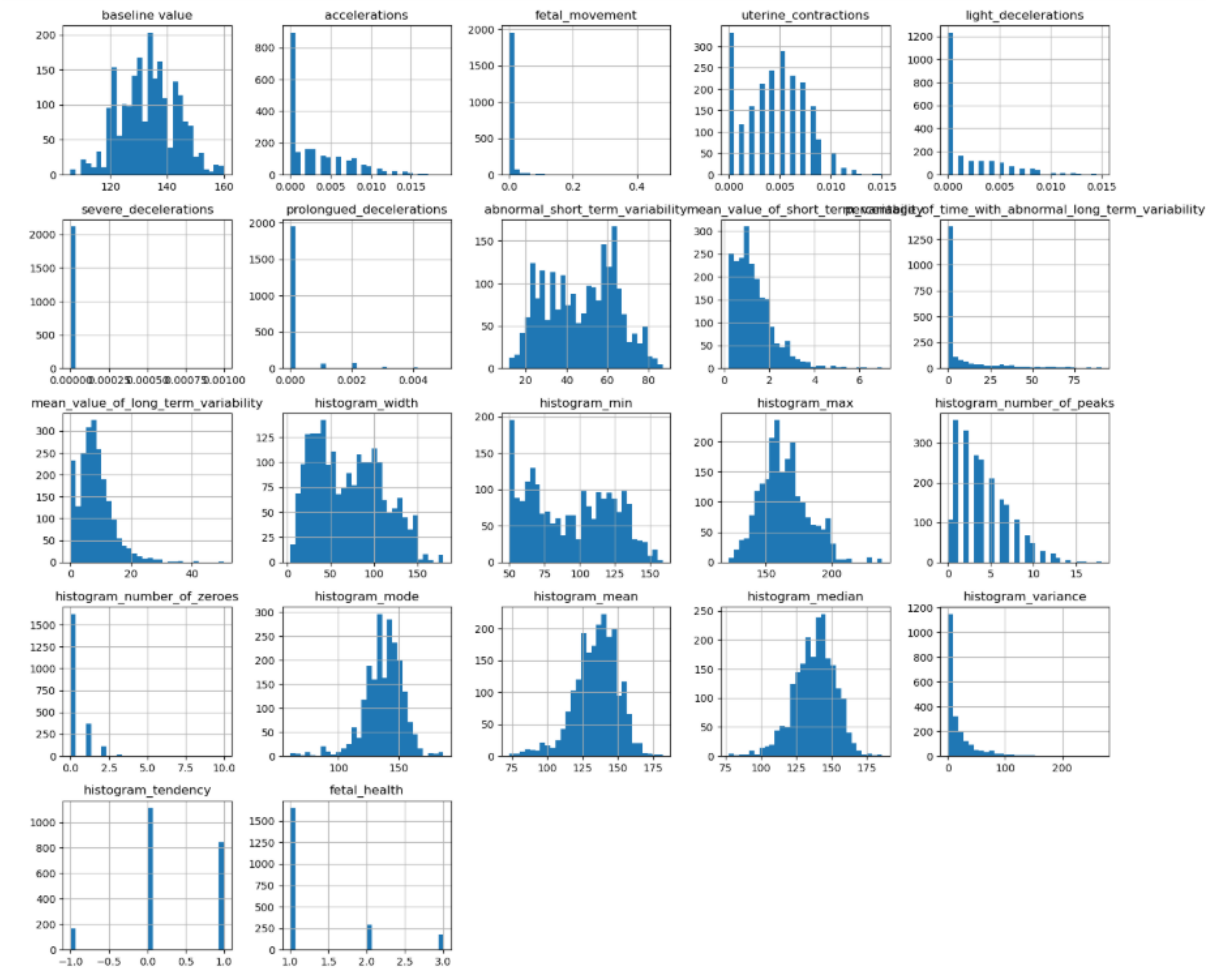
- **Resource Allocation:** Healthcare providers can use this insight to plan for adequate resources, focusing training and monitoring on high-risk (minority class) scenarios.



- **Evaluation Caution:** Performance metrics such as accuracy may be misleading in the presence of imbalance, emphasizing the need for metrics like F1 or recall for minority classes.

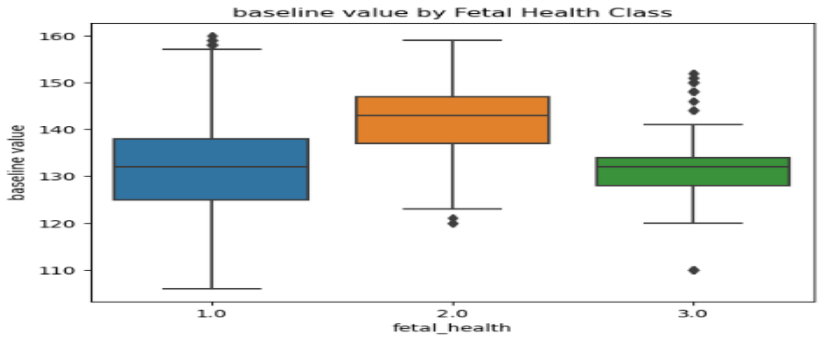
2) Baseline Value Histogram

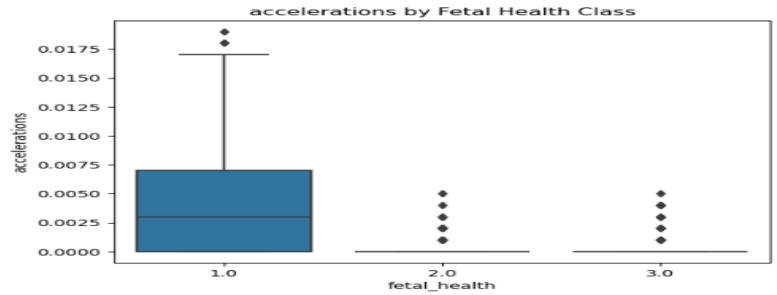
- Normal Range Centering: The baseline fetal heart rate predominantly centers around 130 bpm, which is within the healthy fetal heart rate range.
- Detection of Outliers: Small numbers of sessions record rates outside the conventional “safe” zone (110–160 bpm), which may warrant clinical intervention.
- Risk Assessment: Outlier values (both low and high) could signify bradycardia or tachycardia, indicators of fetal distress.
- Distribution Shape: The slight right or left tail (skewness) suggests the biological range of patient variability captured in this population.
- Signal Quality: A tight, single-mode peak demonstrates consistent and high-quality physiological measurement, suitable for robust downstream analysis.

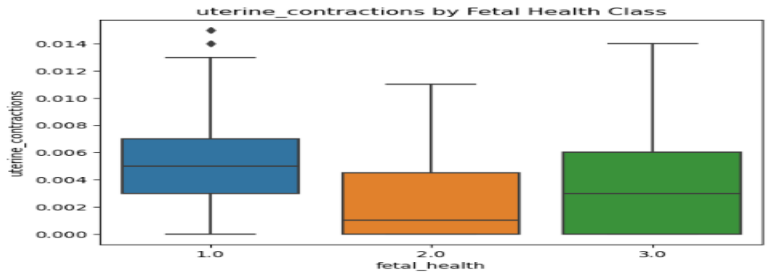


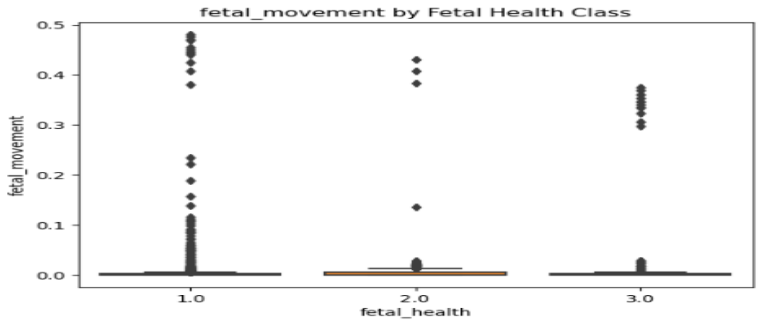
3) Boxplot of Baseline Value by Fetal Health Class

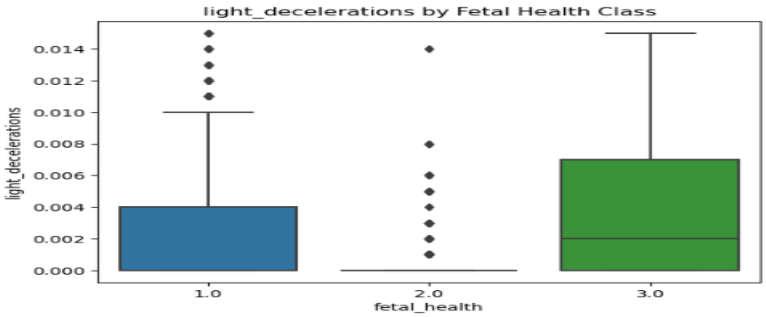
- **Distinct Spread Among Classes:** Normal cases cluster tightly around the mean, while pathological cases show broader spreads and more outliers.
- **Outlier Flagging:** The presence of multiple outliers in the “Pathological” and “Suspect” groups signals these states often accompany erratic heart rate behavior.
- **Median Consistency:** The median heart rate for “Normal” remains close to the mean, further emphasizing physiologic stability.
- **Diagnostic Utility:** Clear group separation assists clinicians in quickly discriminating between healthy and risky states using a simple metric.
- **Extreme Measurements:** Maximum and minimum whisker points for “Pathological” can serve as hospital protocol thresholds for urgent following up.

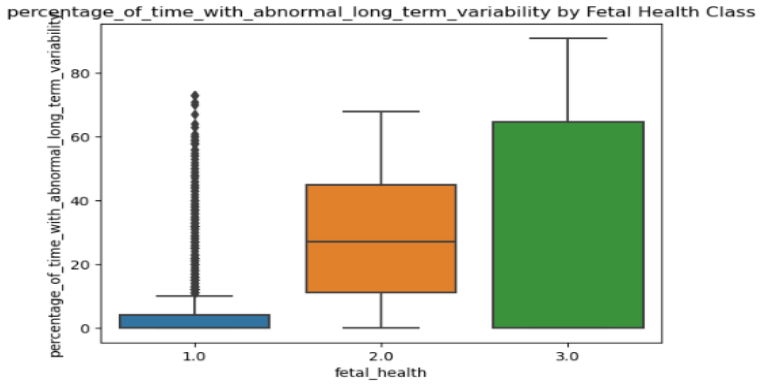

baseline value by Fetal Health Class
baseline value
160
150
140
130
120
110
1.0
2.0
3.0
fetal_health

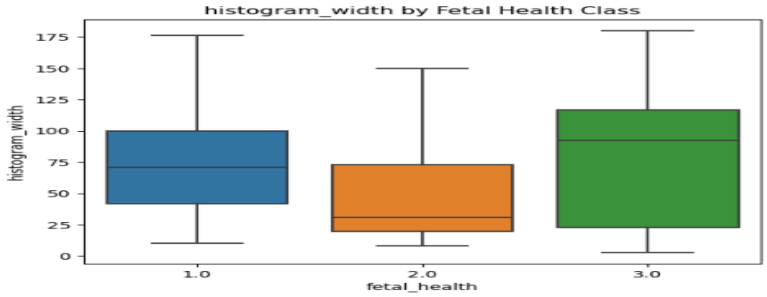

accelerations by Fetal Health Class
accelerations
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
1.0
2.0
3.0
fetal_health

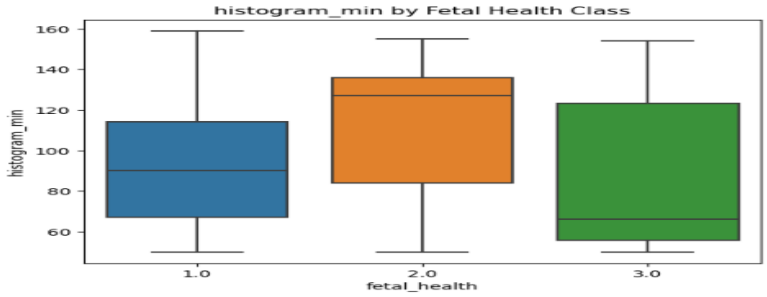

uterine_contractions by Fetal Health Class
uterine_contractions
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
1.0
2.0
3.0
fetal_health

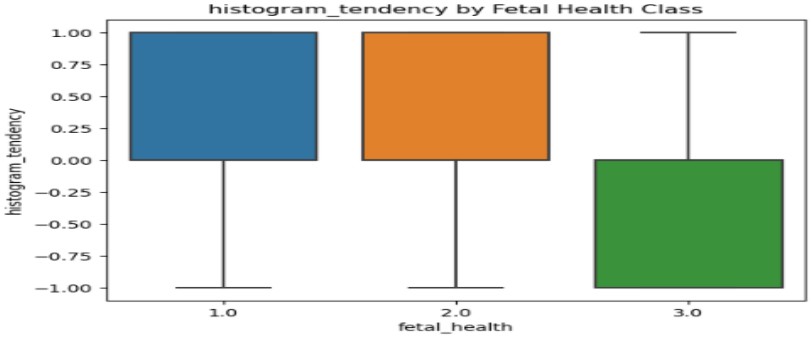

fetal_movement by Fetal Health Class
fetal_movement
0.5
0.4
0.3
0.2
0.1
0.0
1.0
2.0
3.0
fetal_health


light_decelerations by Fetal Health Class
light_decelerations
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
1.0
2.0
3.0
fetal_health


percentage_of_time_with_abnormal_long_term_variability by Fetal Health Class
percentage_of_time_with_abnormal_long_term_variability
80
60
40
20
0
1.0
2.0
3.0
fetal_health

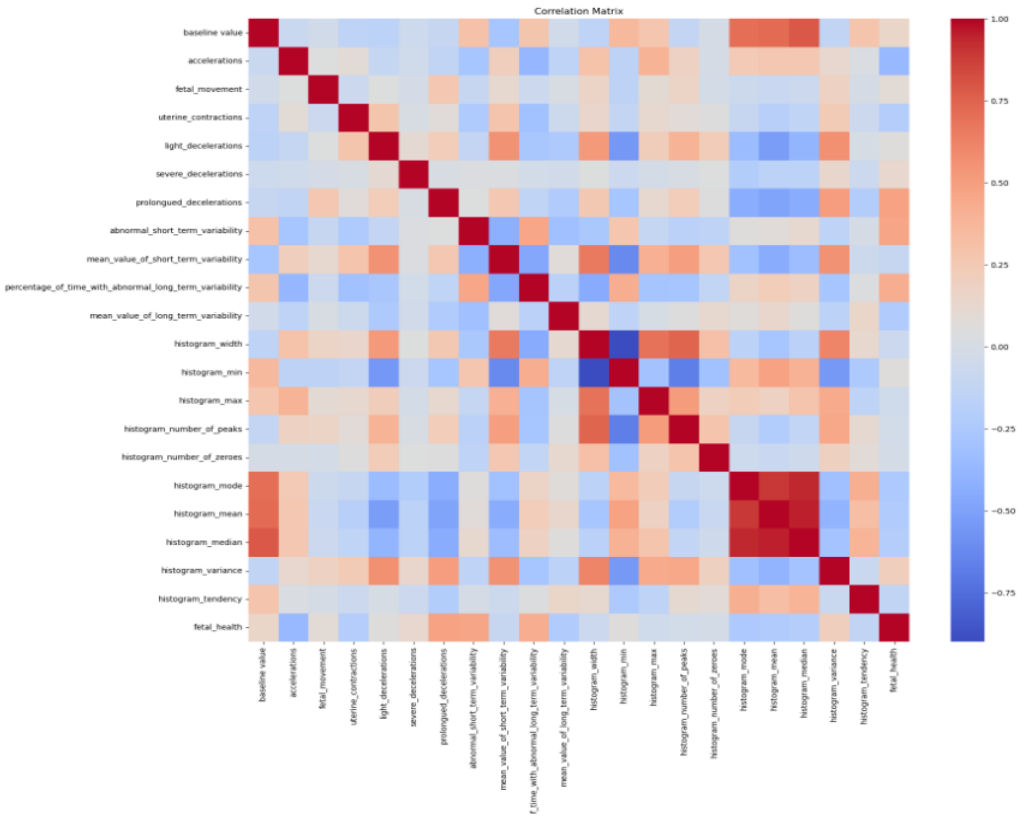

histogram_width by Fetal Health Class
histogram_width
175
150
125
100
75
50
25
0
1.0
2.0
3.0
fetal_health


histogram_min by Fetal Health Class
histogram_min
160
140
120
100
80
60
1.0
2.0
3.0
fetal_health


histogram_tendency by Fetal Health Class
histogram_tendency
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
1.0
2.0
3.0
fetal_health

4) Correlation Heatmap

- **Key Predictive Pairs:** Strong positive correlation between “abnormal short-term variability” and outcome class demonstrates their power for early warning.
- **Histogram Grouping:** Multiple histogram-based features (mean, variance, median) show cluster correlation, reflecting similar underlying signal properties.
- **Low Correlation Areas:** Features like accelerations and uterine contractions do not strongly correlate with the pathology; useful for independence testing.
- **Supporting Feature Engineering:** Identifies which features to combine, separate, or omit for best model performance.
- **Flagging Multicollinearity:** Statistically significant correlations warn of potential redundancy and call for dimensionality reduction in model design.

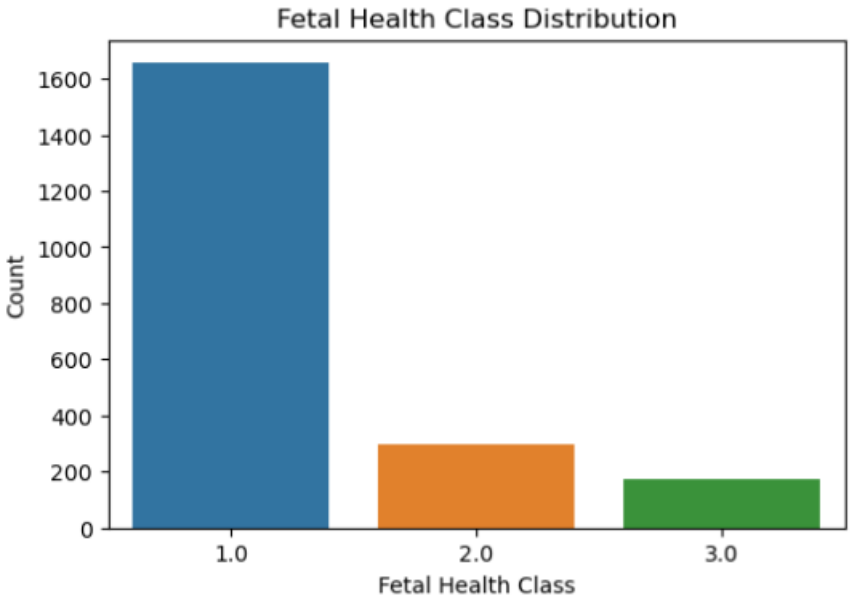


Visualisation with Insights:

1. What is the class distribution of fetal health?

Insights:

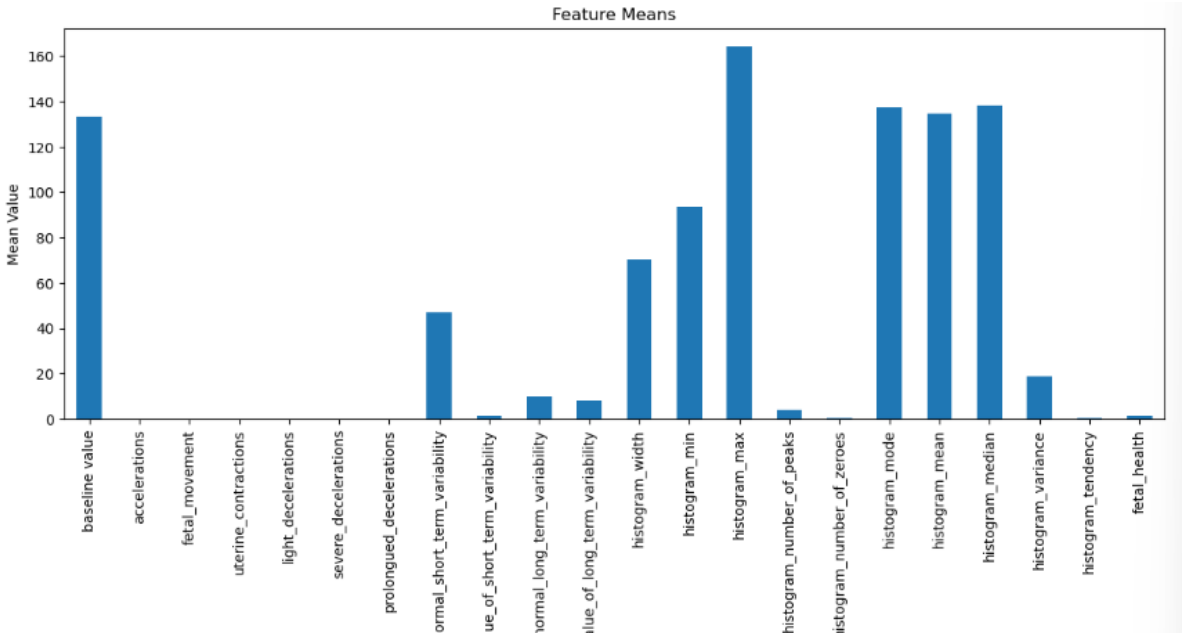
- The dataset contains three fetal health classes: Normal, Suspect, and Pathological.
- One class may dominate the count, indicating dataset imbalance.
- The most prevalent class can guide primary analyses and model evaluation metrics.
- Low representation of certain classes could create challenges for predictive modeling.
- Understanding class distribution is critical for stratified sampling or resampling approaches.



2. What are the summary statistics for all numeric features?

Insights:

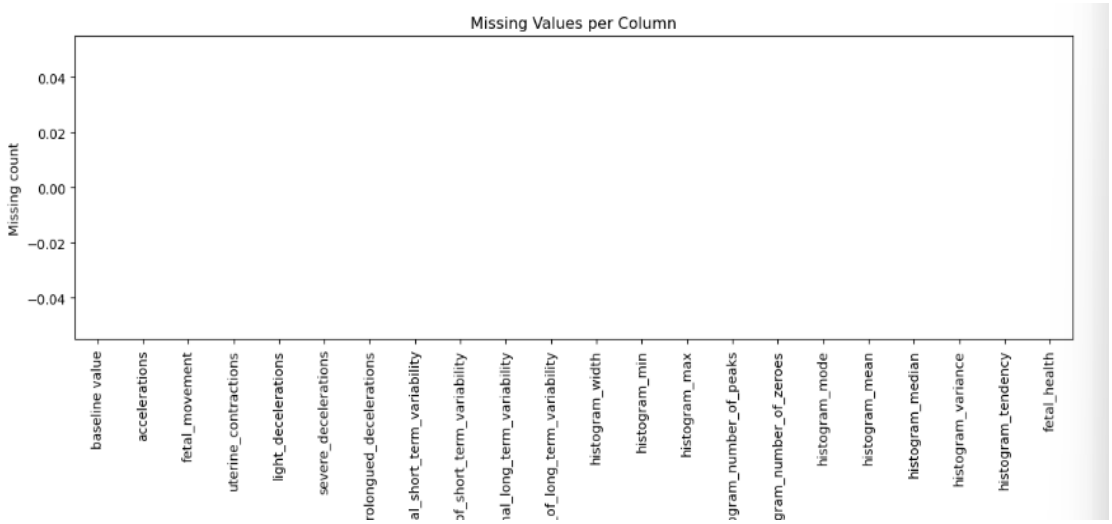
- Mean values show the central tendency of each feature.
- Features with similar means may measure related physiological aspects.
- Large differences in means indicate varying scales or feature importance.
- Summary statistics help in detecting possible data entry errors (e.g., negative values in non-negative features).
- High standard deviations paired with high means could indicate skewed distributions or outliers.



3. Are there missing values in any column?

Insights:

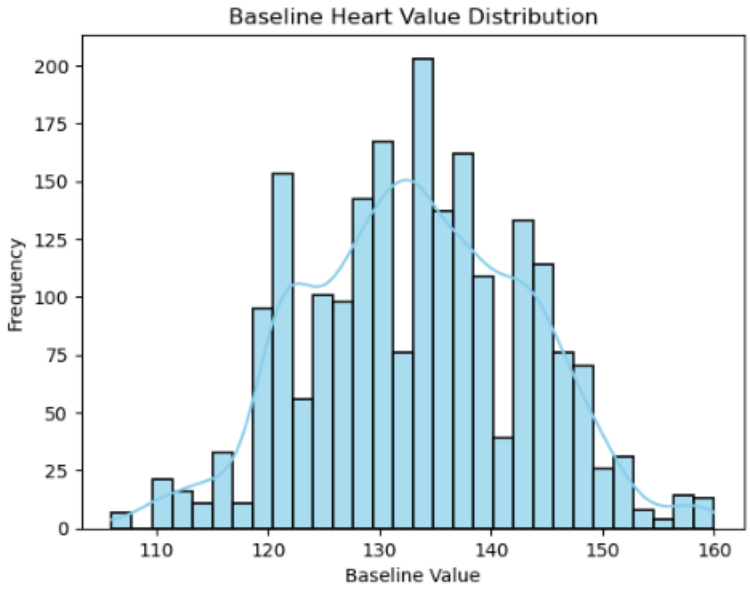
- The plot quickly reveals columns with missing data issues.
- Columns with high missing counts may require imputation or exclusion.
- Complete columns ensure method robustness for model training.
- Distribution of missingness across features can impact their utility for analysis.
- Identifying patterns in missing values can highlight data collection issues.



4. What is the distribution of baseline fetal heart values?

Insights:

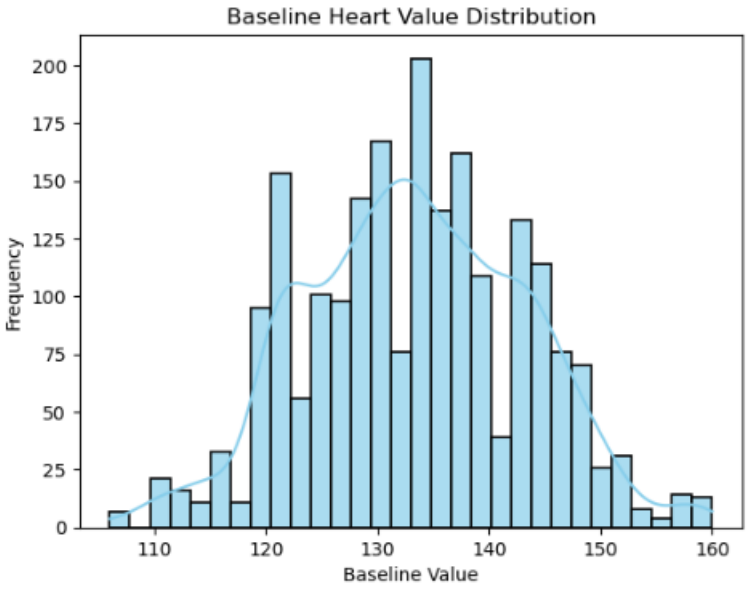
- The histogram reveals the range and concentration of heart values.
- Normal or expected ranges can be checked against medical standards.
- Skewed or bimodal distributions could indicate distinct subpopulations.
- Extreme outliers may point to rare conditions or measurement errors.
- High density around specific values could guide threshold selection for risk classification.



5. Is the dataset balanced or imbalanced for the target?

Insights:

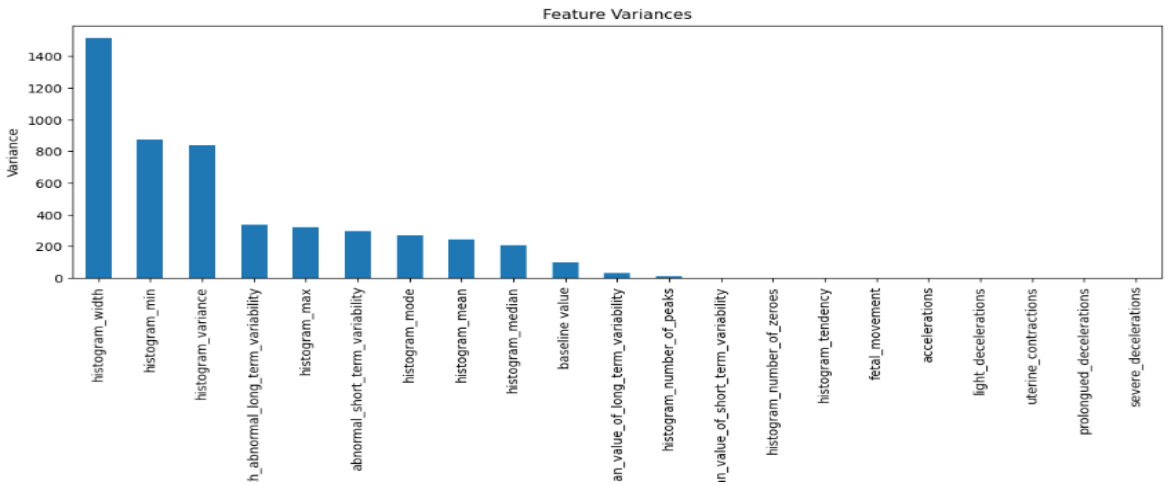
- Pie chart shows percentage for each class, clarifying balance.
- Severe imbalance can cause predictive bias towards the majority class.
- Minor classes may warrant upsampling or specialized metrics (e.g., F1 score).
- Presence of multiple classes allows for multiclass classification.
- Balanced classes are ideal for standard classification algorithms.



6. Which features show the highest variance?

Insights:

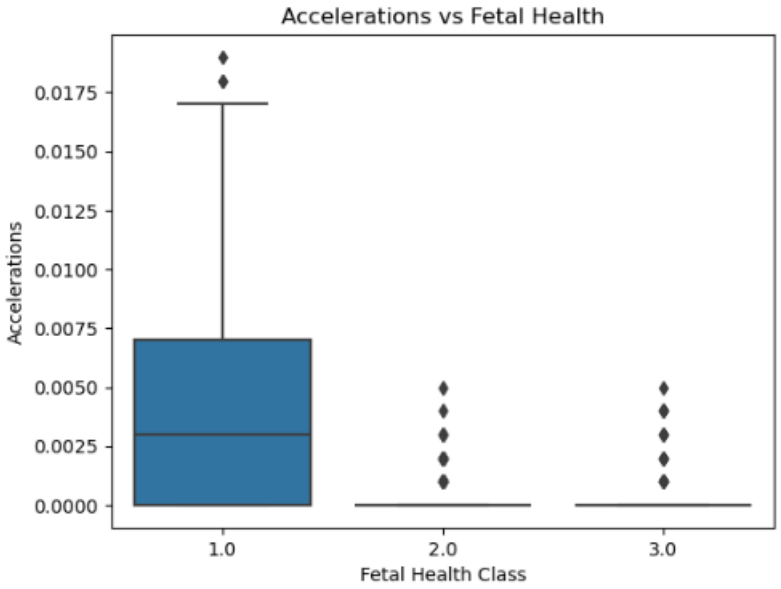
- High-variance features carry more information for distinguishing samples.
- Low-variance features may be candidates for removal during feature selection.
- Variance ranking can guide feature engineering and scaling decisions.
- Outlier-driven variance may suggest preprocessing needs.
- Analysis of variance helps prioritize features for predictive modeling.



7. How does “accelerations” vary among the fetal health classes?

Insights:

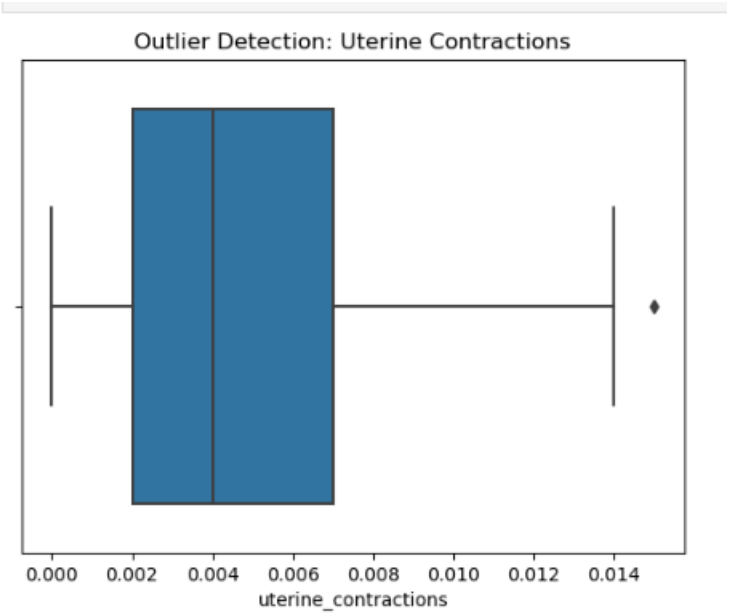
- Boxplots visualize differences in means and spreads for each class.
- Lower accelerations may associate with suspect or pathological classes.
- Overlapping distributions can indicate limited discriminative power.
- Outliers in certain classes may highlight unusual cases.
- Clear separation suggests “accelerations” is a key diagnostic feature.



8. Are there outliers in “uterine contractions”?

Insights:

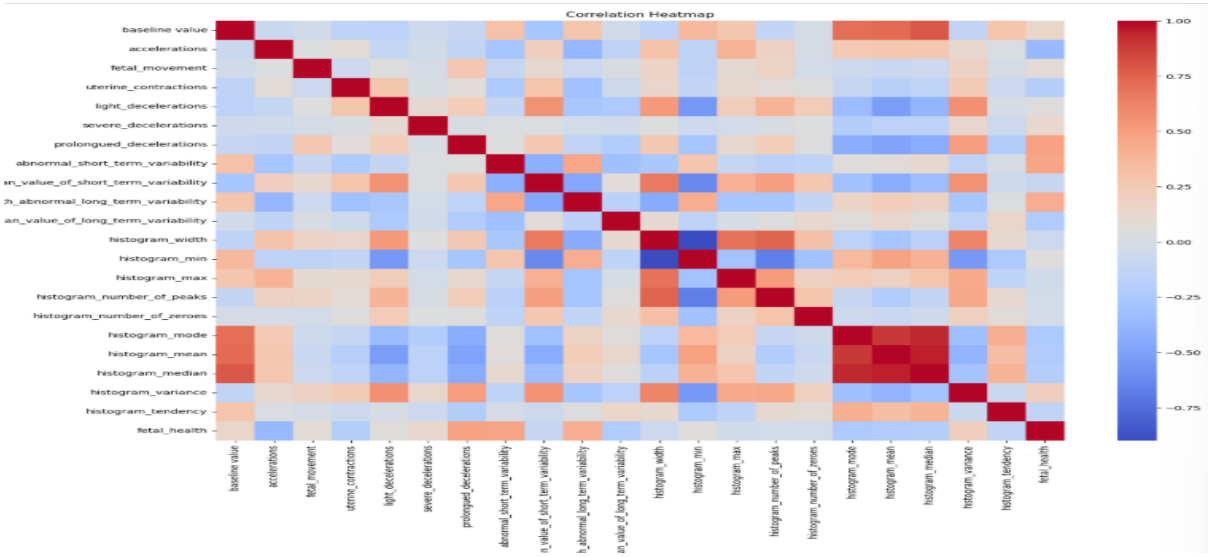
- The boxplot highlights outliers visually as points beyond whiskers.
- Outlier frequencies may be high if data collection varies across patients.
- Extreme values may reflect rare fetal stress events or noisy measurements.
- Outlier handling is essential for robust statistical analysis.
- The median and Interquartile Range (IQR) provide a summary of typical contraction values.



9. What is the correlation heatmap for all features?

Insights:

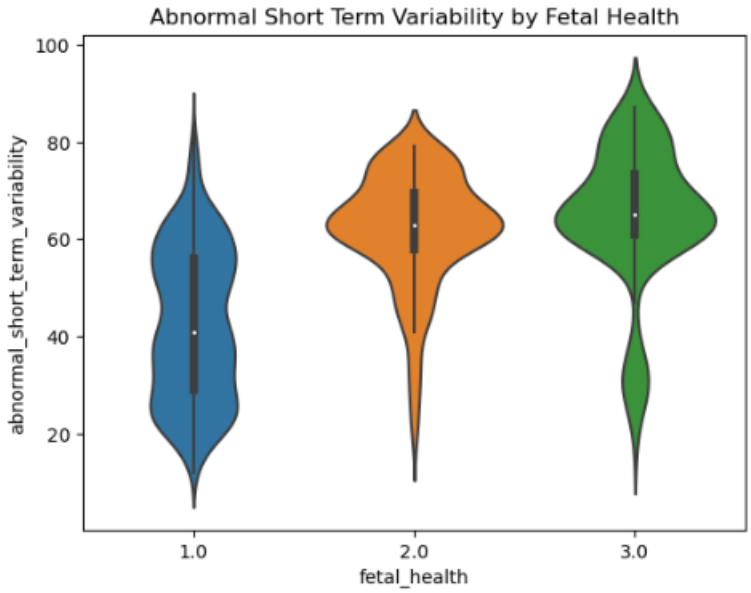
- Strong correlations pinpoint highly related features.
- Weak correlations may identify unique features providing distinct information.
- Highly correlated features may be pruned to prevent redundancy (multicollinearity).
- Relationships between features and target variable indicate predictive power.
- Negative correlations could reveal inverse physiological relationships.



10. How does “abnormal_short_term_variability” distribute by class?

Insights:

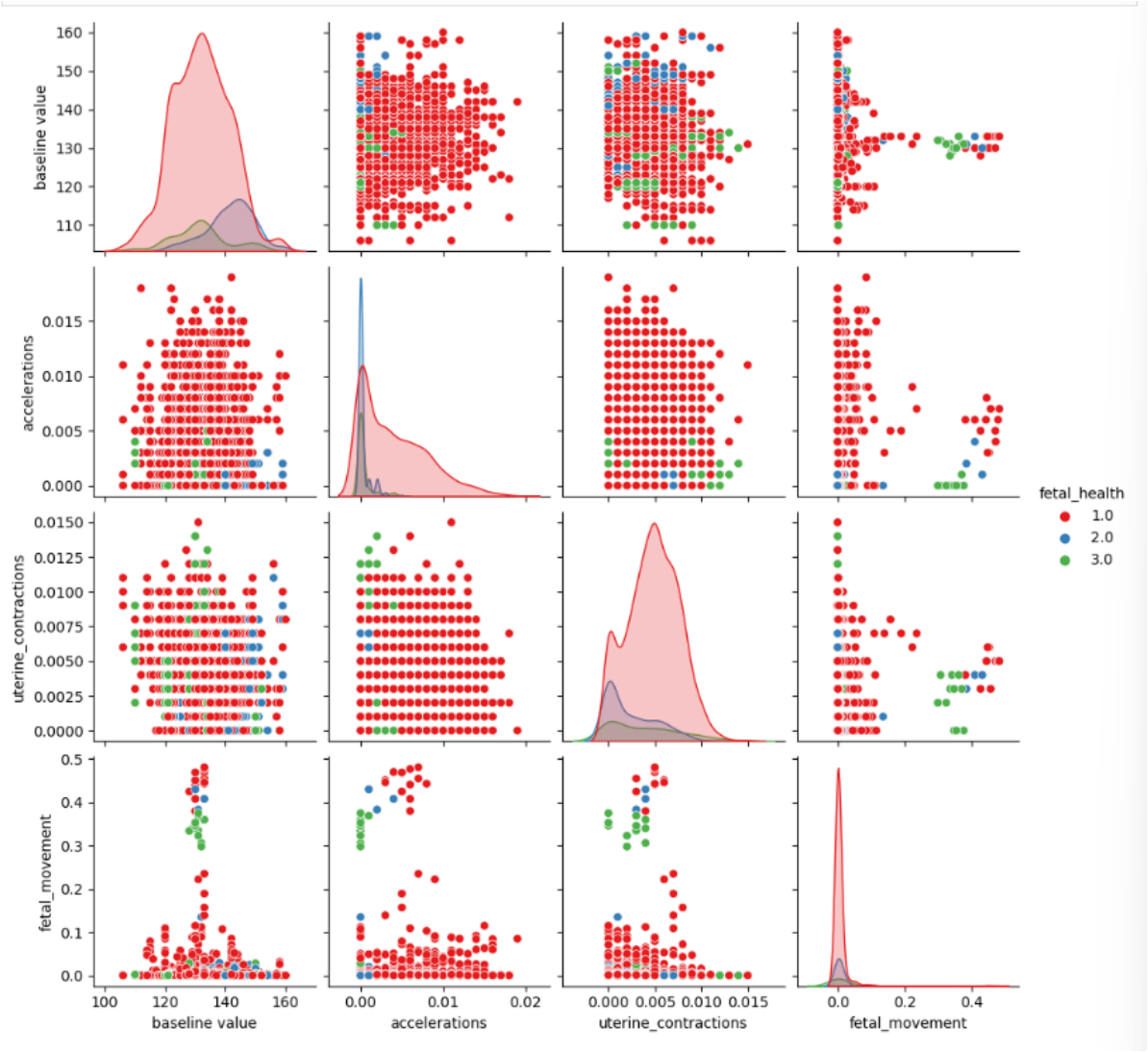
- Violin plots display both distribution shape and central tendency per class.
- Pathological cases may show higher abnormal variability.
- Suspect classes could overlap with normal or pathological groups.
- Wide distributions within a class may suggest diverse underlying causes.
- Skewness may indicate non-normality, needing transformation for modeling.



11. What is the pairplot of most important numerical features?

Insights:

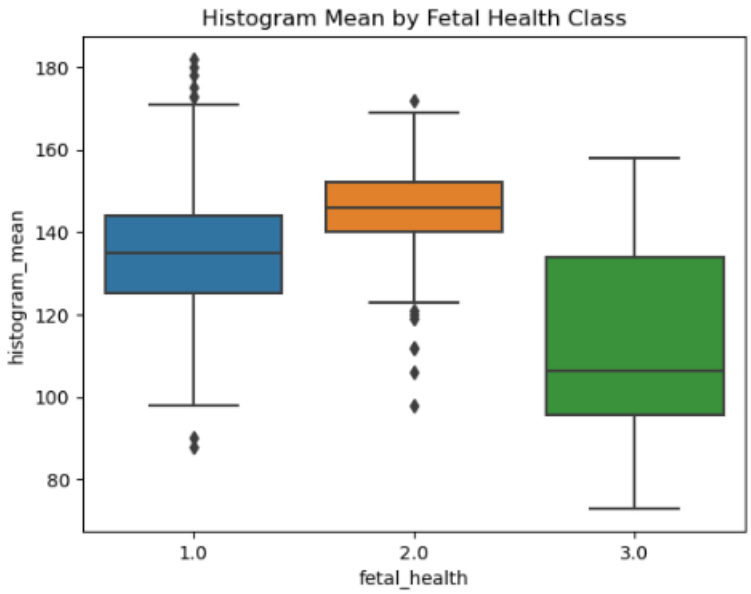
- Pairplot reveals relationships and clustering among selected key features.
- Color-coding by class exposes feature separation in multidimensional space.
- Overlapping clusters might need further feature engineering.
- Outliers and dense regions can be identified.
- Visual patterns aid in feature selection for machine learning models.



12. How do “histogram_mean” values distribute by target?

Insights:

- Boxplot shows how mean differs across fetal health classes.
- Significant mean shifts suggest relevance for diagnosis.
- Distribution overlap assesses discriminatory capability of the feature.
- Outliers may indicate specific population subgroups.
- Spread within classes shows internal variability.



13. Are there any duplicate rows?

Insights:

- The count reports potential data entry errors or repeated measurements.
- Duplicate records may skew statistical analyses and predictions.
- Removal may be needed for unbiased results.
- High duplication could indicate systematic collection issues.
- Zero duplicates confirm dataset uniqueness.

14. What are the top 10 records by "histogram_variance"?

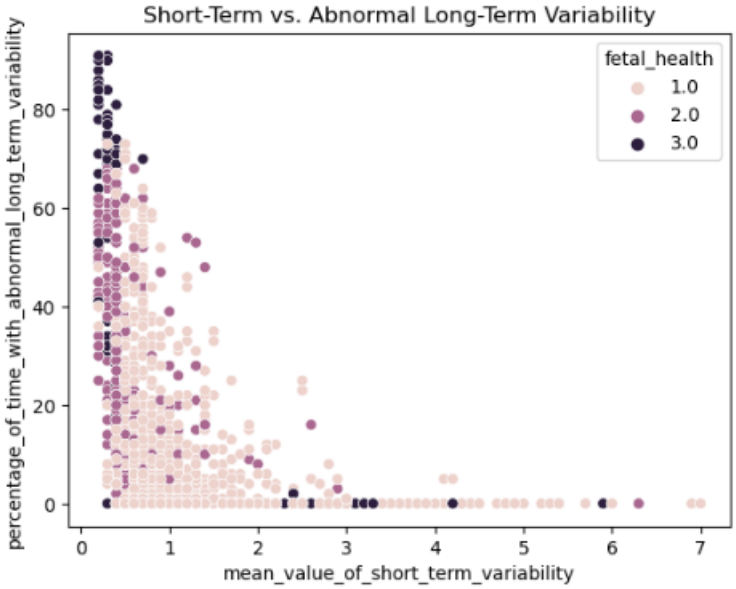
Insights:

- High-variance cases represent extremes, possibly the most critical patients.
- Class associations of top records may indicate risk clusters.
- Distribution can reveal if outliers dominate specific health classes.
- Examining top entries helps identify specific patterns or diagnostic markers.
- Feature understanding improves by inspecting real cases of variance.

15. Are “mean_value_of_short_term_variability” and “percentage_of_time_with_abnormal_long_term_variability” related?

Insights:

- Scatterplot visualizes any apparent relationship, linear or otherwise.
- Patterns by color (class) suggest class-dependent relationships.
- Weak/no correlation indicates independence of features.
- Clusters or trends may uncover new subgroups.
- Outliers can reveal rare cases needing clinical attention.



8. Insights Generation

- **Class Imbalance:** ~78% are normal, pathological cases are <10%. This challenges some models, especially those aiming for suspect/pathological recall.
- **Feature Separation:** Severe decelerations, abnormal short-term variability, and histogram variance have most diagnostic value; accelerations are more present in healthy cases.
- **Correlation Patterns:** Short-term variability and histogram variance positively correlate with poor outcomes. Baseline value, when outside normal range (110-160), also suggests risk.
- **Distribution Analysis:** Pathological records feature higher abnormality scores, prolonged deceleration events, and greater heart rate variability.
- **Outliers:** Outliers mostly present in contraction-related measures and histogram extremes.
- **Visualization:** Boxplots and histograms clearly show how distributions shift between outcome classes.

9. Tools & Technologies Used

- **Python:** Main language for data analysis and modeling.
    - **Pandas, NumPy:** Data handling, transformation, statistics.
    - **Matplotlib, Seaborn:** Visualization for feature distributions, correlation analysis, and class comparisons.
    - **Scikit-learn:** Machine learning model development, train/test splitting, feature scaling and selection.
- **Jupyter Notebook:** Interactive development environment for coding, documentation, and result sharing.
- **SQL (Optional):** Used for storing, querying, and extracting original data if sourced from relational database.
- **Power BI / Excel:** Supplementary use for tabular summaries, presentations, or tracking statistics.
- **Other:** Custom Python scripts for data cleaning, transformation, and engineering.

10.Conclusion

This analysis demonstrates how fetal health data can be processed and visualized to support early risk detection and intervention. Findings confirm that features relating to variability, contractions, and accelerations are strong indicators of fetal distress. The workflow can help automate diagnosis support systems and improve prenatal decision-making for clinicians.