Case Study: Exploratory Data Analysis on Fetal Health Dataset

Objective:

The goal of this case study is to conduct a comprehensive exploratory data analysis (EDA) of the dataset to uncover patterns, relationships, and insights that can assist in understanding factors influencing fetal health.

Dataset Description:

The dataset contains several features related to fetal monitoring, such as baseline values, accelerations, decelerations, and histogram metrics. The target variable fetal_health categorizes fetal health into three groups:

1. Normal (1.0)
2. Suspect (2.0)
3. Pathological (3.0)

Attributes:

- **baseline value**: Baseline fetal heart rate.
- **accelerations**: Number of accelerations per second.
- **fetal_movement**: Movements of the fetus.
- **uterine_contractions**: Number of uterine contractions.
- **light_decelerations**, **severe_decelerations**, **prolongued_decelerations**: Different types of decelerations.
- **histogram metrics**: Metrics derived from the heart rate histogram, including mean, variance, mode, etc.
- **fetal_health**: Target variable indicating fetal health.

Tasks:

Data Overview

1. Load the dataset and display its structure (e.g., number of rows and columns, column names).
2. Identify and handle any missing values in the dataset.

Descriptive Analysis

1. Provide statistical summaries for all numerical columns (mean, median, variance, etc.).
2. Perform univariate bivariate and multivariate analysis
3. Check for missing values and determine their distribution.
4. Analyze the distribution of the target variable fetal_health. Create plots for various analysis.
5. Calculate the summary statistics for all numerical columns.
6. Analyze the distribution of the target variable fetal_health. Visualize the counts of each category.
7. Compute and visualize the correlation matrix. Highlight the top three features with the strongest correlation with fetal_health.

Correlation and Feature Analysis

4. Analyze the variance of key features (e.g., baseline value, uterine_contractions) to identify the most variable metrics.
5. Plot the distribution of baseline value across each fetal_health category.
6. Identify features that have strong correlations with fetal_health. Visualize the correlation matrix.
7. Explore relationships between key features, such as baseline value, accelerations, and uterine_contractions. Use scatter plots or other relevant charts.

Insights and Trends

7. Calculate the average mean_value_of_short_term_variability for each fetal_health category.
8. Compare the mean and median values of percentage_of_time_with_abnormal_long_term_variability between normal and pathological cases.
9. Examine the trend in uterine_contractions and prolongued_decelerations for different fetal health conditions.
10. Examine the distribution of features such as baseline value and mean_value_of_short_term_variability. Plot histograms for these columns.
11. Compare average values of important features (e.g., uterine_contractions, light_decelerations) across the different fetal health categories.

Outliers and Data Quality

10. Use boxplots to identify outliers in histogram_mean, baseline value, and histogram_variance. Discuss how these outliers might affect the analysis.
11. Identify the features with potential data quality issues or inconsistencies.

Visualizations

12. Create a scatter plot to explore the relationship between accelerations and uterine_contractions, color-coded by fetal_health.
13. Generate a pairplot to visualize the relationships among key features grouped by fetal_health.
14. Plot histograms for baseline value and mean_value_of_short_term_variability to analyze their distributions.
15. Identify outliers in columns like histogram_mean or baseline value using boxplots.
16. Investigate which features have the highest variability and their potential impact on fetal_health.
17. Create a pairplot to show how features vary by fetal_health.
18. Visualize the relationship between percentage_of_time_with_abnormal_long_term_variability and the target variable using bar or line charts.

Classification Readiness

15. Evaluate if the target variable fetal_health is balanced. If not, suggest approaches to handle imbalance for machine learning tasks.
16. Highlight key features that might be most predictive of fetal health.

Outcome:

- Develop a deeper understanding of the factors influencing fetal health.
- Identify significant trends, patterns, and relationships in the data.
- Prepare the dataset for predictive modeling by highlighting relevant features and handling potential issues.

Deliverables:

1. **EDA Report**: A detailed analysis of the dataset, including visualizations, summaries, and insights.

Summarize your findings in a concise report, including key insights about fetal health.

Propose actionable steps for further analysis or feature engineering based on your EDA.

Deliverables:

1. A Jupyter Notebook or script with all your code and visualizations.
2. A written report summarizing your key findings, supported by visuals.
3. Recommendations for preparing the dataset for predictive modeling, if applicable.