Project Methodology – INFO-B529

Predicting Fetal Health from Cardiotocogram Data - A Machine Learning Approach Towards Reducing Child and Maternal Mortality

Group members – (Lakshmi Aparna Valiveti, Suneetha Naidu Mekala)

AIM: This study aims to utilize machine learning techniques to develop a predictive model that accurately assesses fetal health status based on Cardiotocogram (CTG) data. Through the analysis and classification of CTG records into three distinct categories: Normal, Suspect, and Pathological, our objective is to create a robust and reliable tool for healthcare professionals. By leveraging machine learning algorithms, we aim to provide an efficient and effective means of early detection of potential fetal health complications, thereby contributing to the reduction of child and maternal mortality rates, particularly in low-resource settings.

PURPOSE:

Fetal health monitoring is critical throughout pregnancy and childbirth, as it enables early detection of complications and timely interventions to safeguard the well-being of both the baby and the mother. Techniques like Cardiotocography (CTG) provide essential insights into fetal well-being, allowing healthcare professionals to intervene promptly, if necessary, thus reducing the risk of outcomes like birth asphyxia or neonatal complications. Ultimately, ensuring optimal fetal health through vigilant monitoring plays a pivotal role in reducing child and maternal mortality rates and promoting the overall well-being of families and communities (Grivell et al., 2015). The primary purpose is to develop a robust predictive model capable of accurately assessing fetal well-being based on CTG records. By categorizing CTG data into Normal, Suspect, and Pathological classes, the aim is to equip healthcare professionals with a reliable tool for early identification of potential fetal health risks, thereby facilitating timely interventions to mitigate adverse outcomes. The project has the potential to significantly improve mother and child health outcomes, especially in resource-constrained areas with limited access to advanced medical equipment and expertise.

Furthermore, this project is consistent with worldwide efforts to accomplish the United Nations Sustainable Development Goals for reducing child, maternal mortality rates. By enhancing the capacity to detect and address fetal health complications early in pregnancy, this research aims to contribute to the overarching goal of ensuring safer pregnancies and childbirth experiences for women worldwide. Through the integration of machine learning methodologies into prenatal care, the ultimate purpose is to foster positive health outcomes for both mothers and newborns, ultimately striving towards the vision of a healthier and more equitable world for all (Hasan et al., 2019).

DATA DESCRIPTION:

1. **Dataset:** The dataset has been collected from Kaggle, https://www.kaggle.com/datasets/andrewmvd/fetal-health-classification/data

2. **Variables:** There are 22 attributes with 2127 rows in the dataset

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 min, Histogram max, Histogram number of peaks, Histogram number of zeroes, Histogram mode, Histogram mean, Histogram median, Histogram variance, Histogram tendency, Fetal health.

METHODOLOGY:

DATA EXPLORATION & PREPROCESSING:

This involves a thorough examination of the dataset's structure, identifying variable types (categorical or numerical), understanding potential significance, and discerning the relationships with target variable. This phase also includes handling missing data, treating outliers, and performing feature scaling or normalization as needed.

a) Missing data will be handled either by removing them or by substituting them with mean or media and with mode in case of categorical variables.
b) Detection and treatment of outliers using inter-quartile range (IQR) or by using Z- score. Identifying and addressing outliers is crucial for ensuring the integrity, accuracy, and interpretability of data analyses and predictive models, ultimately leading to more robust and reliable insights and decision-making.
c) Standardization or Normalization: Standardize numerical features to a similar scale to prevent any bias in the model towards features with larger magnitudes. Common methods include Min-Max scaling for Normalization or Z-score for Standardization.

DATA VISUALIZATION:

Summary statistics: The terms mean, median, range, variance, and standard deviation characterize the variables central tendency and dispersion.

Exploratory Data Analysis (EDA): This step serves as a crucial initial step in understanding the underlying structure and characteristics of a dataset.

1. ***Box Plots:*** Utilize box plots to visualize the central tendency, data distribution, and potential outliers for each numerical attribute, facilitating key feature comparisons between instances of fetal health categories (Normal, Suspect, Pathological).

2. ***Histograms:*** Employ histograms to evaluate data trends and distribution patterns of the continuous variables. This visualization will illustrate the distribution of attributes such as Baseline value, Accelerations, and others across the dataset.
3. ***Scatter Plots***: Create scatter plots to gain insights into associations between variables, allowing examination of correlations and trends. For instance, investigate relationships like the impact of Fetal movement or Uterine contractions on fetal health status.
4. ***Bar Plots:*** Visualize categorical variables such as the prevalence of different fetal health categories (Normal, Suspect, Pathological) or the amount of time with anomalous long-term variability allows for a clear comparison of distributions and their relevance to fetal health.
5. ***Heat Maps:*** Create heat maps to emphasize the intensity and direction of correlations between variables, which will aid in the identification of interdependence and significant relationships. This will assist in understanding relationships among attributes like Fetal health and other physiological parameters, potentially revealing critical factors influencing fetal health status.

MACHINE LEARNING MODELS:

Based on our project purpose, our dataset has independent and dependent variables

Independent variables: 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 min, Histogram max, Histogram number of peaks, Histogram number of zeroes, Histogram mode, Histogram mean, Histogram median, Histogram variance, Histogram tendency,

Dependent/ Outcome/ Predicting Variable: fetal health

Among all these variables based on the results from the correlation matrix we would perform feature engineering to build the prediction model.

1. ***Logistic Regression***:
   a) Logistic regression uses a logistic function to predict probabilities, making it especially effective for binary classification problems.
2. b) It's an appropriate model for binary or multi-class classification problems. As our outcome variable is in multi-class, we included logistic regression in our analysis.

3. ***Decision Trees***
   a) Decision trees divide data into subsets based on the value of input features, and make judgments at each node based on a feature. This process continues until the model achieves a significant level of prediction accuracy or meets a stopping criterion.
   b) Given the multi-class nature of our fetal health outcome variable, decision trees are particularly beneficial for our project as they can handle complex, nonlinear relationships between features and outcomes.

4. ***Random Forest:***
   a) Random Forest is an ensemble approach that constructs several decision trees during training time and outputs the class that is the average of the classes in the individual trees.
   b) In case of our project on predicting fetal health, Random Forest offers a robust solution by aggregating the predictions of multiple decision trees, thereby significantly reducing the risk of overfitting that might occur with a single decision tree.

5. ***Gradient Boost Machines (GBM):***
   a) Gradient Boosting Machines develop an ensemble of weak prediction models, often decision trees, in a stage-wise method, optimizing for a loss function and correcting errors caused by previously trained trees.
   B) In our project, GBM is exceptionally valuable due to its ability to iteratively minimize errors, offering precise predictions even in the presence of complex and nonlinear relationships between features.

6. ***Support Vector Machines (SVM):***
   a) SVM creates a hyperplane in a high-dimensional space to distinguish between classes. It searches for the largest marginal hyperplane that best splits the dataset into classes.
   B) For our project, SVM's strength lies in its ability to handle the high-dimensional nature of our dataset, effectively distinguishing between the classes of fetal health.

7. ***K-Nearest Neighbors (K-NN):***
   a) K-NN categorizes a data point according on how its neighbors are classed. It aggregates the classes of the closest samples and guesses the class based on the majority, using feature similarity to generate predictions.
   b) In our project on predicting fetal health, K-NN could be particularly effective due to its ability to make predictions based on the similarity of cardiotocogram measurements. Based on the scenario that fetal health conditions exhibit patterns or clusters within the data, K-NN can leverage these patterns to accurately classify fetal health status.

MODEL EVALUATION METRICS

a. Accuracy
b. Classification Report involving Precision, Recall, and F1-Score
c. AUC-ROC Curve
d. Validation
e. Confusion Matrix

Some of these or a combination of these metrics would be used to evaluate model performance, balancing overall accuracy with the nuanced needs of medical decision-making.

INTERPRETATION OR FEATURE IMPORTANCE:

Coefficients in Linear Models: For models like logistic regression, the coefficients can indicate the importance and direction of the relationship between each feature and the outcome.

Tree-based Feature Importance: Decision trees and ensemble methods like Random Forest and Gradient Boosting Machines offer straightforward metrics to evaluate feature importance based on how often a feature is used to split the data and how much it decreases the impurity.

Permutation Feature Importance: This technique involves randomly shuffling a single feature and measuring how much the model's performance decreases. A significant drop indicates high importance.

SHAP Values: SHAP (SHapley Additive exPlanations) values offer a way to understand the contribution of each feature to every prediction, considering interaction effects. This method provides both global and local interpretability.

EXPECTED RESULTS:

- We expect that the application of machine learning models to cardiotocogram data will significantly enhance the ability to accurately predict fetal health status.

- By leveraging the predictive power of algorithms such as Random Forest, Gradient Boosting Machines (GBM), and Support Vector Machines (SVM), among others, we aim to achieve a high level of accuracy, precision, and recall in distinguishing between normal, suspect, and pathologic fetal conditions.

- Through the careful analysis and interpretation of cardiotocogram data, this project aims to set a precedent for the integration of machine learning techniques in obstetrics, enhancing the standard of care and supporting clinicians in making informed decisions for both expectant mothers and their babies.

REFERENCES

Grivell, R. M., Alfirevic, Z., Gyte, G. M., & Devane, D. (2015). Antenatal cardiotocography for fetal assessment. Cochrane Database of Systematic Reviews, 9(9). https://doi.org/10.1002/14651858.cd007863.pub4

Hoodbhoy, Z., Noman, M., Shafique, A., Nasim, A., Chowdhury, D., & Hasan, B. (2019). Use of Machine Learning Algorithms for Prediction of Fetal Risk using Cardiotocographic Data. International journal of applied & basic medical research, 9(4), 226–230. https://doi.org/10.4103/ijabmr.IJABMR_370_18