**Final Project Proposal**

1. **Project Title**

Fetal Health Classification

1. **Names(Last name, First name)**

Bai Xiaoxue

Jiang Bei

Liu Yuchen

1. **Problem statement**

The United Nations highlights the reduction of child and maternal mortality as a critical goal, aiming to lower under-5 deaths to fewer than 25 per 1,000 live births by 2030. In low-resource settings, where the majority of maternal deaths occur, accessible diagnostic tools like Cardiotocograms (CTGs) can play a pivotal role. CTGs, which assess fetal health by monitoring heart rates, movements, and uterine contractions, offer a cost-effective strategy for early detection of potential complications. This project focuses on the classification of fetal health using CTG data to enhance prenatal care and prevent child and maternal mortality, addressing a significant gap in healthcare in underserved areas.

1. **General approach**

The approach to classifying fetal health from CTG data involves several key steps, beginning with data preprocessing to handle missing values and normalize the data, ensuring that the scale of different features does not bias the model. Following this, the dataset is split into training and test sets to evaluate the performance of the models accurately.

A variety of machine learning models, including Logistic Regression, Decision Trees, Gradient Boosting, Random Forest, KNN and SVM, are then trained on the dataset. Pipelines are set up for each model to streamline the process of fitting and predictions. The performance of these models is assessed using cross-validation to ensure that the results are reliable and not dependent on a particular split of the data. Then choose one with best performance as the baseline model to proceed the analysis.

Given the imbalanced nature of the dataset, with a higher proportion of normal cases compared to suspect or pathological ones, one challenge will be ensuring that the models do not become biased towards predicting the majority class. Techniques such as oversampling the minority class or adjusting class weights might be necessary to address this issue.

Another challenge is the selection of appropriate features and model hyperparameters. This is addressed through exploratory data analysis to understand the distribution and correlation of features and by using GridSearchCV to find the optimal model parameters that yield the best cross-validation score.

The final step involves evaluating the chosen model's performance on the test set, using metrics like accuracy, precision, recall, and the F1 score, alongside the interpretation of confusion matrices to understand the model's strengths and weaknesses in classifying fetal health.

1. **Data**

The dataset, sourced from Kaggle, comprising 2126 records from Cardiotocogram (CTG) exams forms the foundation of our project to classify fetal health into three categories: Normal, Suspect, and Pathological. Each record in this dataset encapsulates various features extracted from CTG exams, which have been meticulously classified by three expert obstetricians into the aforementioned categories. The inclusion of expert classifications provides a robust ground truth for training and validating our classification models.

The dataset is ideal for our project due to its comprehensive range of fetal health indicators, such as heart rate, uterine contractions, and movements. This allows for a detailed analysis of fetal well-being. Its direct relevance to reducing child and maternal mortality through early intervention aligns with our goal of improving prenatal care, particularly in resource-limited settings. Additionally, the dataset's size and variety offer a robust foundation for training diverse and effective machine learning models, enhancing the project's applicability and potential impact.

1. **Measure of success**

Success for this project will be measured through several key performance indicators aligned with the objectives of improving prenatal care and reducing child and maternal mortality:

* Model Accuracy: A high accuracy indicates that the model can correctly classify the fetal health status in most cases, which is crucial for effective prenatal care. An accuracy rate above 90% would be considered a success.
* F1 Score: A high F1 score for each of the fetal health categories, especially for the 'Pathological' class, would indicate a robust model.
* Area Under the Curve: An AUC close to 1.0 would indicate an excellent model performance, with a score above 0.9 being considered highly successful in accurately identifying the fetal health status while minimizing the rate of false positives and negatives.

1. **Milestones and deliverables**
2. Data Preprocessing and Exploration: This involves loading the dataset, performing initial exploratory data analysis, visualizing missing values , and understanding the distribution of features and the target variable. Tools and techniques like histograms, count plots, and correlation heatmaps will be utilized for this purpose.
3. Feature Scaling and Data Preparation: Standardize the features in the dataset using techniques such as Standard Scaling to ensure that the model's performance is not biased by the relative scales of the features. Split the data into training and testing sets to facilitate model training and evaluation.
4. Model Selection and Pipeline Setup: Establish pipelines for different classifiers including Logistic Regression, Decision Tree, Gradient Boosting Classifier, Random Forest, KNN and SVC. This step ensures a structured approach to applying and evaluating multiple machine learning models.
5. Model Training and Cross-Validation: Train the models using the training dataset and perform cross-validation to evaluate their performance. This step aims to identify the most effective model based on the cross-validation scores.
6. Hyperparameter Tuning: For the most promising model, perform hyperparameter tuning using techniques like GridSearchCV to find the optimal set of parameters that yield the best performance.
7. Final Report Evaluation: Evaluate the final model using the testing dataset. Key Performance Indicators such as accuracy score, F1-score, and confusion matrix will be used to assess the model's performance.
8. Project Report: Compile a comprehensive report detailing the methodology, implementation details and results.