Fetal Health Project Plan (Chen Yoffe)

Project objective:

The objective of this project is to design a machine learning model that can predict fetal health using cardiotocograph (CTG) interpretation

Cardiotocography (CTG) is a technique used to monitor the fetal heartbeat and uterine contractions during pregnancy and labor. The machine used to perform the monitoring is called a cardiotocograph.

The dataset:

The dataset includes 2126 records of 21 features extracted from Cardiotocogram exams, which were then classified by three expert obstetricians into 3 classes:

* Normal (=1)
* Suspect (=2)
* Pathological (=3)

this is the outcome variable (*e.g* the predicted value).

The project will be done using Python and mainly the following Packages: pandas, NumPy, matplotlib.pyplot, seaborn, and scikit-learn.

Project flow:

1. EDA of the data

* Statistical analysis for all variables: mean and standard deviation, median and IQR (25%-75%) min, max, count.
* Outlier detection - If the values are definitely wrong, convert to NA or take out the observation.  
  Otherwise, analyze the influence in the presence and absence of the outliers on the final model.
* Check for missing values – there are no missing values in the dataset chosen for my project
* Use graphs to get an insight into the data:  
  Pair-plot of the variable to get information on variable interaction
* Correlation matrix – Done to explore correlations between variables and identify redundant variables.
* Categorical variables of interest can be displayed using bar-plots and numeric variables by histogram. Interaction of variables can be displayed using box-plots when testing numeric vs categorical variables and scatter-plots to test numeric vs numeric interactions.

1. Feature selection- Table one (to choose variables with a significant p-value (< 0.05)) can be used in addition to other modeling methods like RandomForest, and SVM, and then choosing the most voted variables.
2. If the data is imbalanced regarding the outcome variable (if one of the categories in the “fetal health” variable is found in less than 10% of the observations), balancing needs to be done. Balancing the data can be done and tested by several methods (Synthetic Minority Oversampling Technique, Undersampling, Oversampling). The best-performing method should be used only on the training dataset.

## Normalize the numeric variables to be on the same scale.

1. Divide the data to train and test (Dev can be used as well), this can be done by using the sklearn package and the train\_test\_split function.
2. Model selection: supervised multi-classification models such as Decision Trees, Naive Bayes, RandomForest, SVM, KNN and Gradient Boosting can be used in this case. Each model will be trained, fit and then evaluated.
3. Evaluation of the model can be done using AUC, Accuracy, Recall and Precision. The AUC should be used as the most accountable matric since it integrates both the recall and the specificity matrices.
4. Choose the best performing model and adjust its hyperparameters