EAS509 STATISTICAL LEARNING & DATA MINING - II
FINAL PROJECT REPORT

1.1 Motivation/Problem Statement:

Cardiotocograms are widely used for assessment of fetal wellbeing and has been the mainstay of fetal monitoring for over 50 years. The fetal heart rate fluctuations and their temporal correlation with uterine contractions are captured by cardiotocography (CTG). The goal of obtaining a cardiotocography analysis is to identify infants who may be anemic (hypoxic) in order to direct further evaluations of fetal welfare or decide whether the baby must be delivered through caesarean section or assisted vaginal birth. The diagnostic features were measured and automatically processed from 2126 Fetal Cardiotocograms (CTGs). Classification and Clustering methods were used to classify the samples into (Normal-1, Suspect-2, Pathologic-3) to provide adequate and appropriate healthcare.

Dataset (Source: https://archive.ics.uci.edu/ml/datasets/cardiotocography)

The dataset consists of 2300 instances of fetal cardiotocogram features, and all the 11 features were used to cluster the patients into categories.

Methods/Models:

I. Data Cleaning/Exploratory Data Analysis

We deployed 4 different classification models on the dataset. First the dataset was cleaned, scaled, and divided into a 70-30 train: test split ratio. No two pairs of features showed any high correlation, and hence we used all 11 features for the analysis.

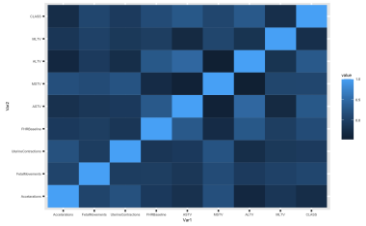
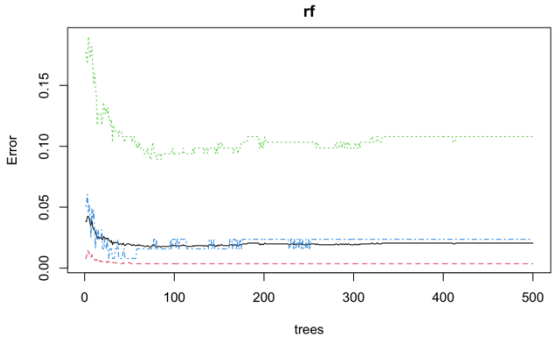
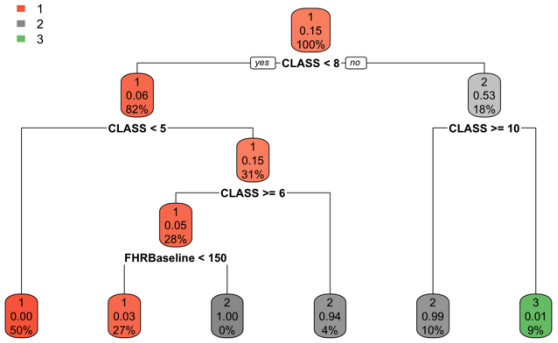


Figure 1.1: Data Cleaning

II. Classification

As the next step, the models were trained with k-Nearest Neighbor, Random Forest, Naïve bayes and Decision Tree classifiers. The naïve bayes classifier achieved the highest accuracy of 99.93%, and a very less error rate (depicted in the figure below), and hence we chose this model to predict the Fetal State Class (FSC) that the patient falls under.





The Decision Tree model provided a closely high value of accuracy of 99.36% for the training set, and the functioning of the algorithm is depicted in the figure above. As the number of observations of 'Fatal' in our dataset is less, the model derives only one case where the patient is classified into class 3.

III. Clustering

Using all the features, this time including the Fetal State Class (FSC), we perform K-Means Clustering, Hierarchical Clustering, to cluster our data into different groups of patients, which would help us identify closely-related patients based on how similar their features are.

K-means Clustering:

We performed k-means clustering using various methods like elbow, silhouette, and gap statistics. The elbow method was used to to determine the optimal number of clusters into which the data may be clustered. Silhouette analysis was used to study the separation distance between the resulting clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like number of clusters visually. The Gap statistic standardized the graph of log(Wk), where Wk is the within-cluster dispersion, by comparing it to its expectation under an appropriate null reference distribution of the data. The optimal number of clusters we obtained from the k-means clustering using all the 3 methods were 2. (k = 2). The cluster visualization using the K-means clustering is as follows:

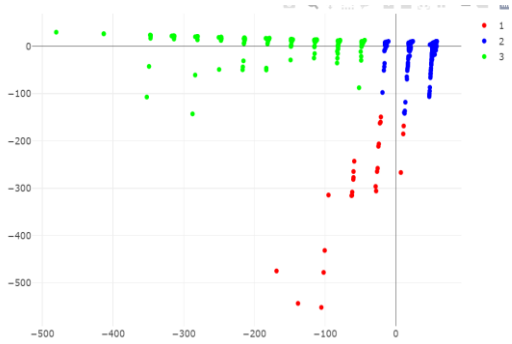


Figure 1.2 K-means Clustering Visualization

Hierarchical Clustering:

Each data point was treated as a singleton cluster in hierarchical clustering, which then merges clusters one at a time until only one is left. A dendrogram was plotted for visualizing the hierarchical clustering. When using complete-link we combined the two clusters with the smallest maximum pairwise distance. When using single-link we combined the two clusters whose two nearest members are separated by the least distance at each step. Average-link clustering, strikes a balance between single-link clustering's propensity to generate extended chains that don't fit the intuitive idea of clusters as compact, spherical objects and complete-link clustering's sensitivity to outliers.

Conclusions:

With the help of K-means clustering, we could see that the accuracy is higher when k value is 2 which means that the output is rendered with noise since, the k value is smaller.

GitHub link for code: https://github.com/suriyabadrinath/eas509prj

Members: Abhiram Siddoju (50442313), Gowtham Kumar (50442715), Kaviyaa Vasudevan (50443082), Suriya Badrinath (50442121),