MET CS 677 Sprint 2 – Term Project Summary

Sushant Khot

04/29/2021

Topic Selected: Fetal Health Classification

Introduction:

I decided to choose this dataset having experienced the birth of my premature kid who is doing great today. I wanted to examine and study the dataset to understand what analysis we can do on the fetal health data.

Tasks performed:

- Examine which features have strong correlation with Class label.
- Perform Classification Analysis and build models using :

    1. KNN Classifier (find optimal k and re-run to get the final accuracy)
    2. Logistic Regression classifier
    3. Naive Bayesian
    4. Decision Tree
    5. Random Forest

I calculated and discussed Performance Metrics for these classifiers by preparing Confusion Matrix to look at how our prediction models perform.

I have Split Data into Training and Testing to verify the models built (50/50).

I also tried to visualize the dataset and look for features that can help for classification, outliers, correlation etc.

Some Questions that I tried to answer at the end of Project and analysis:

1. What features in different Visualizations show us trends and help classify a fetus.

2. Compare and build insights on different features and their correlation with the Class Label.

3. Which Model best predicts the Health of a fetus?

4. Compare all classifiers listed above and discuss our findings using confusion matrix.

Data:

The “Fetal Health” dataset contains 2126 records of features extracted from Cardiotocogram exams, which were then classified by three expert obstetricians into 3 classes:

Normal, Suspect and Pathological.

I have combined "Suspect" and "Pathological" classes into one class called "Abnormal" and Normal will stay as "Normal".

Conclusion:

We started with analyzing our Fetal health dataset and it was pretty clean dataset to begin with and we did not have to do any pre-processing activities.

Some features did have outliers, but we decided not to remove them as the medical data entry errors would be minimum and we went ahead without removing any outliers.

We created various visualizations to help support our analysis and gather more insights on the features present in the dataset.

For e.g. Count plot gave us an idea on distribution of our dataset based on class label.

Correlation Matrix provided us the top most correlated and non-correlated features with Fetal Health

Box plots provided us with insights on data distribution within each feature.

Regression plots provided us with correlation and distribution of features around the 2 class labels "Normal" and "Abnormal" fetal health and fetal movement data.

We split our dataset 50/50 into training and testing and ran through multiple classifiers on the data.

We documented the statistics in terms of classifier Accuracy to correctly classify the Fetal health and discussed the confusion matrix.

Finally based on the confusion matrix we concluded that based on Overall Accuracy, KNN (k = 3), Random Forest and Decision tree seemed to have performed well.

Random Forest is able to classify more "Normal" class fetal health correctly.

KNN (k = 3) and Decision Tree have predicted more "Abnormal" class Fetal health correctly.