**MACHINE LEARNING REPORT ON MEDICAL DATASET**

The project consists of implementing decision tress for classification and regression on a medical dataset

**Objective of the project**

This project will implement a classification Decision Tree algorithm on a Fetal cardiology medical dataset. It will classify the Fetal case (NSP) and predict to which of the following categories will the fetal case belong to:

N=normal (1);

S=Suspect (2);

P=Pathologic (3)

Important: this project will use a black-box approach making use of other’s packages to analyze the information. The black-box approach has been chosen for its simplicity but needs to be noted that this approach entails a great peril of not understanding what happens within the function and end up with wrong results.

The dataset is made 14 decision variables and 1 class variable.

· BPM (Beat per minutes)

· APC (Accelerations per second)

· FMPS (Fetal movement per second)

· UCPS (Uterine contractions per second)

· DLPS (Light declaration per second)

· SDPS (Severe declaration per second)

· PDPS (Prolonged declaration per second)

· ASTV (% of abnormal short term Variability)

· MSTV (Mean of short term Variability)

· ALTV (% of abnormal long term Variability)

· MLTV (Mean of long term Variability)

· Width (Width of FHR Histogram)

· Min (Min Width of FHR Histogram)

· Max (Max Width of FHR Histogram)

· NSP (Fetal State Class codeN=normal (1); S=Suspect (2); P=Pathologic (3))

**Data Cleaning**

The dataset is already cleaned and it is ready for performing the decision tree.

Performing classification Decision Tree

this section will show how to perform a classification decision tree on the medical dataset

Let’s start by exploring the data by calculating the 5 numbers summary of each variable.

Text

Description automatically generated

We can see that the NPS which is a class variable is an INT, however It needs to be a converted into categorical variable to predict.

Text

Description automatically generated

In order to perform the decision trees we will need to split the dataset in 2. The training dataset and the test dataset for validation. In this instance, the weight of the training is 80% and the validation is 20%.



Start the model training

Diagram

Description automatically generated

Confusion matrix

Text

Description automatically generated

Calculate classification accuracy = 0.8063725

classification error = 0.1606519

# Conclusion

A decision tree is an excellent tool for classification problems and for its simplicity that allows the user to interpret the information very quickly. However, the decision tree models suffer the variance, correlation, and overfitting problems.

The implementation of the classification decision trees has achieved its purpose of building a satisfactory model to classify Fetal State. However, the results could be improved to make a sound model able to be implemented. In that sense, a bigger dataset will be needed to train more the model.

**References**

<https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/>

<https://en.wikipedia.org/wiki/Decision_tree>