Roland T. Killian

Mais202

MAIS 202 - PROJECT DELIVERABLE 1

I will be using the NIH GLP13534 dataset, which contains over 100,000 pieces of information on Methylated DNA (https://www.ncbi.nlm.nih.gov/gds/?term=GPL13534[Accession]). I am choosing this data because it is vast and usable. It has the methylation profile of various samples as well as malignancy and tissue type formatted in text files that can be easily converted to CSV's. The Dataset contains methylation data as well as a boolean of malignant or not. Some specific data also contains percent malignancy. I will start by looking at brain tissue and hope to determine the percent tumor in a given sample. If all goes well, I hope to branch out and create an algorithm that can determine tissue type as well as malignancy.

I will start by gathering data of brain tissue and their corresponding Methylation profile. The Methylation profile is very large, so I will start with a small number of samples, and using feature extraction techniques will obtain the most important features.

From this data set, I want to start by writing a classification algorithm that can separate inputted samples to malignant and non malignant. From here, I might move to classification of tissue type as well (instead of just brain).

Currently I am undecided on my approach to solving this problem. I might use logistic regression, naive Bayes, or support vector machines. I will also look into neural networks and see if it would be feasible.

I will make sure to report the mean squared error if investigating percent tumor, and percent accuracy for the classification approach.

I will try to implement my project into a website generated with hugo, gatsby, or react. I will display graphs and explain the project as well as provide a user input, where they can input a csv of a methylation profile and get an output to test my algorithm themselves.

I am not too sure as to the extent of this project and don't know exactly what my final algorithm will do. At the very least however, it should be able to classify malignant brain tissue from non malignant brain tissue (from non brain tissue).