**Team 4 Proposal: Project 2 – StrokePrediction**

Team Members:

* Danielle Dejean
* Kaidon Kennedy
* Carolyn Scheese
* Harpreet (Monty) Singh

Data Set:

As a team we discussed various options, finally deciding on a health care problem. We decided to see if we could train an algorithm to predict for stroke.

Harpreet identified 2 related data sets from Kaggle.com, which is considered a credible source.

1. <https://www.kaggle.com> Stroke Prediction Data Set which contains eleven clinical features for predicting stroke events. <https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>
2. <https://www.kaggle.com> Framingham heart study dataset <https://www.kaggle.com/datasets/aasheesh200/framingham-heart-study-dataset>

Together we reviewed the data sets. Although both sets could meet the assignment requirements (greater than 1000 records), we determined that set #1 was best for our project because it had very few rows that would need to be dropped (fewer than 200 out of 5000 rows) due to “NA’s”.

We created a Slack Channel named *project-2-group-4* to help us communicate effectively.

We set up a GitHub project (project2\_strokeprediction) with branches.

We created a .ipynb file to clean the data.

We have decided to use a “Supervised ML” model because there is a column that identifies who has had a stroke. We are using classification methods.

Thus far we have been familiarizing ourselves with the dataset and doing some data cleaning.

We plan to use and apply the following: Scikit-learn, Pandas, Python, Matplotlib, Pandas plotting. We plan on creating an interactive streamlit app which users can use to predict their individual risk for stroke.

We will use PPT for our final presentation.