### San José State University
## Computer Science Department

## CS156, Introduction to Artificial Intelligence, Spring 2022

## Term project progress report

## Name: Bernard Tan

## SID: 015215317

Provide the following information for your term project.

1. **AI problem and background:**
Stroke is a condition in which the arteries leading to and inside the brain get blocked. It is the fifth largest cause of death and a major source of disability. When a blood vessel carrying oxygen and nutrients to the brain is either blocked by a clot or breaks, a stroke occurs (or ruptures). There have been other previous solutions to prevent it such as quit smoking, limit alcohol use, keep a moderate weight, and get regular checkups People who suffer from this disease and company who work in medical field want to predict, prevent, and improve the accuracy that leads to the disease by using the help of AI. One of the ways for AI to help would be to test the correspondents in lab with the equipment needed. After receiving the data from the machine, it will be inputted and let the AI predict it. I agree with this approach as we need many data to get the high accuracy result.

2. **Dataset of choice:**
https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset
This dataset contains several information that can help to predict stroke. The data contains 5110 observations with 12 attributes.
I can download the csv file.

3. **Describe the independent variables:**
Id = unique identifier of correspondent (Numeric)
Gender = sex of correspondent (Categorical)
Age = Age of correspondent (Numeric)
Hypertension = Hypertension of correspondent (Numeric)
Heart_disease = Heart disease of correspondent (Numeric)
Ever_married = Marital status of correspondent (Categorical)
Work_type = Work type of correspondent (Categorical)
Residence_type = Residence type of correspondent (Categorical)
Avg_glucose_level = Average glucose level in blood (Numeric)
Bmi = Body mass index (Numeric)
Smoking_status = smoking status of correspondent (Categorical)

4. **Describe the dependent variable:**
Stroke = The result that determine if the correspondent had a stroke or not (Binary Categorical)
Yes, this variable is aimed to get. Based on the variable, this is a classification problem. The number of classes is 2.

5. **Describe the data splits**
I am using the cross-validation, with train: test split of 80:20. I think I will differentiate my split, to have more similar distributed data for the stroke prediction.

6. **Number of training observations:**
4088 Data Points
Class 0: 1990

Class 1: 2098

7. **Number of validation observations:**
   N/A

8. **Number of test observations:**
   1022 Data Points
   Class 0: 476
   Class 1: 546