**Final Project**

**Stroke Prediction Data**

Akila Selvaraj

Masters of Data Science,

Bellevue University

DSC 530 T301

Aug 12,2022

**Statistical Question**

There is a saying that stroke will come to patients with type 2 diabetics.

*People with high glucose level tend to get stroke?*

Above is the statistical question I took for exploratory data analysis. For this analysis I took the stroke prediction dataset available at <https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset> . This dataset has columns such as gender, age, hypertension, heart\_disease, work\_type, residence\_type, bmi, and avg\_glucose\_level. Out of all these columns, I chose stroke as the dependent variable and bmi, age, hypertension, heart\_disease, avg\_glucose\_level as explanatory variables. I chose these columns for exploratory analysis as I felt these attributes have more influence on the stroke.

**Outcome of EDA**

Some of the insights from EDA are

* Patients with stroke are having higher avg\_glucose\_level than 120.
* Usually, people who smokes have more chances of stroke but the chosen dataset doesn’t show much impact of smoking on stroke.
* Stroke has high correlation with age.
* The risk of getting stroke increases as the age increases especially beyond 60.

**Other variables**

In the health dataset, out of eleven columns, I have considered only 5 variables for the analysis. Other variables which are not considered are gender, race and job, which could have impact on stroke. Stroke impact could vary based on the gender of the person. Type of job could also influence the stress level of an individual which in turn could impact the health and could be one of the causes of stroke. Lifestyle and food habits of the people differ in rural and urban places and it could also be one of the independent variables. Regression results and the accuracy of the model could be different if we have included other variables as well into consideration.

**Assumptions**

I have assumed that bmi of person has impact on avg glucose level and those two variables are correlated to each other. But the correlation coefficient shows that the correlation between those is less.

**Challenges faced**

Understanding the p-value and interpreting the results correctly was the difficult and challenging part of hypothesis testing. Specifying and null hypothesis and alternative hypothesis and set the significance level need statistical knowledge and I also believe proper interpretation of statistical evidence is important to take intelligent decisions.