MGT 6203 Group Project Proposal

**TEAM INFORMATION (1 point)**

**Team #:** 30

**Team Members:**

1. Melody Lin; sillyapplepie13579; A auditor with limited programming experience. Currently this is the only class I am enrolled in.
2. Wen Yu; amberloisyu; A data analyst in a bank with programming experience in Python but unfamiliar with R. Currently enrolled in both ISYE 6501 and MHT 6203 on EDX in preparation for the OMSA Fall 2022.
3. Erfan Hajibandeh; erfanhajibandeh; A traffic engineer with a M.Sc. in Transportation Engineering. With limited programming experience. Currently enrolled in ISYE 6501 and MHT 6203 in preparation for the OMSA Fall 2022 intake.
4. Shitong Song; songshts1995; A software engineer living in Seattle. I have some programming experience but am not familiar with R. I am enrolled in both ISYE 6501 and MGT 6203 on EDX.
5. Mengchiao Li; mli101; A risk manager in credit card industry. Using SQL and SAS in the work and moving to Python very soon. Having R experience in ISYE 6501, also enrolled MGT 6203 and CSE 6040 on EDX.

**OBJECTIVE/PROBLEM (5 points)**

**Project Title:** Stroke Prediction

**Background Information on chosen project topic:**

According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths. The selected dataset will be used to predict whether a patient is likely to get a stroke based on the input parameters like gender, age, various diseases, and smoking status. Each row in the data provides relevant information about the patient.

We are all coming from a non-medical background and this is an exciting opportunity for us to apply analytical models in an area outside of our day-to-day work context.

**Problem Statement (clear and concise statement explaining purpose of your analysis and investigation):**

Using the stroke dataset, we are aiming to generate predictive models to predict stroke probability.

**State your Primary Research Question (RQ):**

Is there any strong relation between stroke and the variables in the data set?

**Add some possible Supporting Research Questions (2-4 RQs that support problem statement):**

1. Will marriage increase the risk of getting a stroke?
2. Does stroke have a strong relation with other health indicator factors (such as smoking)?
3. Which predictors are most correlated with the rate of stroke?

**Business Justification:** **(Why is this problem interesting to solve from a business viewpoint? Try to quantify the financial, marketing or operational aspects and implications of this problem, as if you were running a company, non-profit organization, city or government that is encountering this problem.)**

The outcome of our analysis is expected to indicate that stroke is correlated with the health status of individuals, and the following markets will be impacted:

Health Insurance: Increasing the insurance premium for individuals with a high probability of stroke.

Advertisement: Strategy changes in targeting the individuals by the various markets such as health care providers, dietary supplement industry, etc.

Medical institutions: Improving more routine care for individuals who have a high chance of stroke and are susceptible

**DATASET/PLAN FOR DATA (4 points)**

**Data Sources (links, attachments, etc.):**

https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

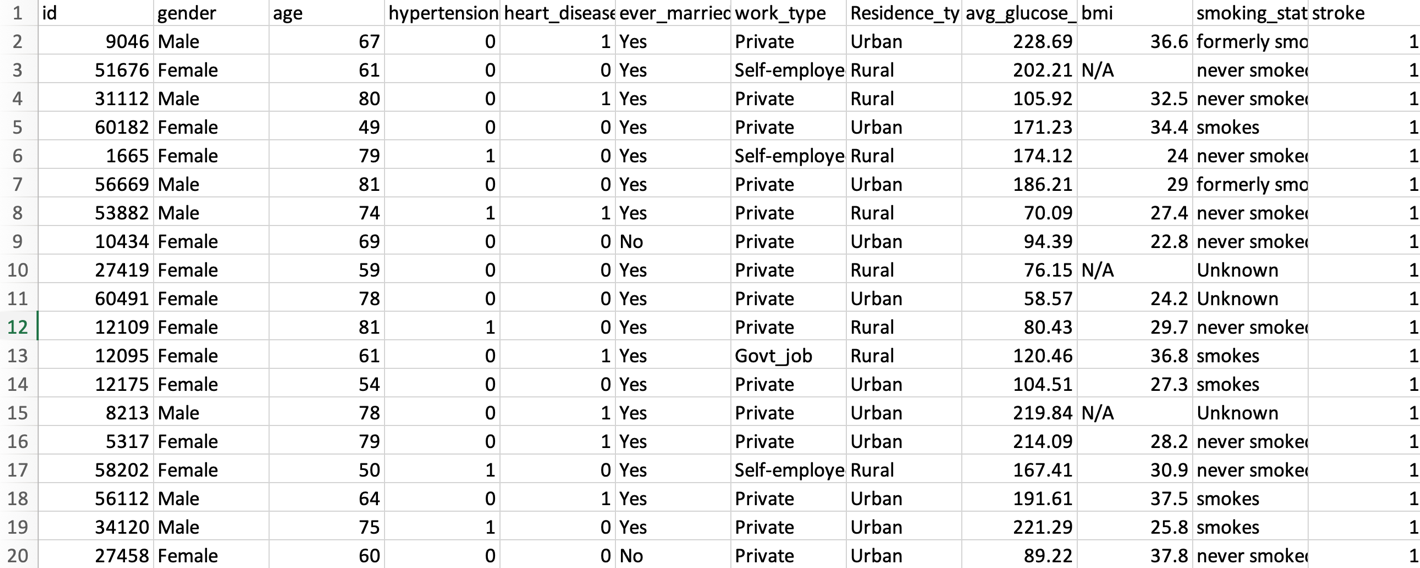
**Data Description (describe each of your data sources, include screenshots of a few rows of data):**

This dataset is used to predict whether a patient is likely to get a stroke based on the input parameters (e.g. gender, age, various diseases, and smoking status). Each row in the data provides relevant information about the patient. This data set has 5111 data points.

* id: unique identifier
* gender: "Male", "Female" or "Other"
* age: age of the patient
* hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
* heart\_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
* ever\_married: "No" or "Yes"
* work\_type: "children", "Govt\_jov", "Never\_worked", "Private" or "Self-employed"
* Residence\_type: "Rural" or "Urban"
* avg\_glucose\_level: average glucose level in blood
* bmi: body mass index
* smoking\_status: "formerly smoked", "never smoked", "smokes" or "Unknown"\*
* stroke: 1 if the patient had a stroke or 0 if not

\*Note: "Unknown" in smoking\_status means that the information is unavailable for this patient

Screenshot of dataset



**Key Variables: (which ones will be considered independent and dependent? Are you going to create new variables?** **What variables do you hypothesize beforehand to be most important?)**

Independent: Gender, Age, Hypertension, Heart Disease, Marriage Status, Work Type, Residence Type, Average Glucose Level, BMI, and Smoking Status.

Dependent: Stroke status

We will create interaction variables, as needed

Hypothesized most expected important parameters: Age, Hypertension, Heart Disease, Smoking Status, Average Glucose Level, BMI

**APPROACH/METHODOLOGY (8 points)**

**Planned Approach (In paragraph(s), describe the approach you will take and what are the models you will try to use? Mention any data transformations that would need to happen. How do you plan to compare your models? How do you plan to train and optimize your model hyper-parameters?))**

We are intending to take the following steps in our analysis:

1. Brief data review to understand the variables and their definition
2. Data cleaning to eliminate the data points that are incomplete or complete the missing data
3. Data exploration analysis to understand the relationship between the independent variables and the dependent variables
4. Introducing dummy variables and data transformation, as needed
5. Splitting the data into a training set (e.g. 80%) and a testing set (e.g. 20%)
6. Using cross-validation to select the optimal parameters and model. Then, train the data using the following models to predict/classify the data:
7. Logistic Regression
8. Random Forest / Decision Tree
9. SVM
10. KNN
11. Evaluating the performance of the model and explaining the relation between the variables
12. Conclusion

**Anticipated Conclusions/Hypothesis (what results do you expect, how will you approach lead you to determining the final conclusion of your analysis) Note: At the end of the project, you do not have to be correct or have acceptable accuracy, the purpose is to walk us through an analysis that gives the reader insight into the conclusion regarding your objective/problem statement**

We are anticipating a meaningful relation between the stroke occurrence probability with health-related variables and age. The mathematical conclusion will be paired with some research facts to support the analysis.

**What business decisions will be impacted by the results of your analysis? What could be some benefits?**

The results of our analysis are mostly expected to benefit healthcare business and insurance strategies. The individuals who are more susceptible to stroke can receive better care from health care institutions and the insurance providers can define their policies for these particular individuals accordingly.

**PROJECT TIMELINE/PLANNING (2 points)**

**Project Timeline/Mention key dates you hope to achieve certain milestones by:**

June 22, 2022: Proposal submission

June 25, 2022: Meeting to divide responsibilities

July 2, 2022: Exploratory data analysis and preparing project progress submission

July 6, 2022: Project progress submission

July 9, 2022: Model development

July 16, 2022: Final project preparation

July 20, 2022: Final project submission

**Appendix (any preliminary figures or charts that you would like to include):**

Not at this time