**Data Science Program Final Project**

**Executive Summary**

The purpose of this study is to bring awareness regarding heart attack and stroke and emphasize on the potential risk factors that play a significant role into the occurrence of these cardiovascular disease.

**Business Objectives**

Using the skills acquired from the data science tutorial, we will attempt to determine the leading key factors that may play a significant role in the occurrence of cardiovascular disease such as stroke and heart attack.

**Background**

Recent studies have shown that stroke is the fifth leading cause of death in United States. Along with heart attack, stroke, and other cardiovascular disease account for about 320 billion in healthcare expenses annually. Knowing the risk factors early enough that are responsible of these chronic cardiovascular diseases can help prevent them from occurring.

**Scope**

Using Python, R and Tableau, we will attempt to display the correlation that exist between heart attack and stroke and the risk key factors underlined in our dataset. To execute this scope, an average working hour of 5 hour per day is planned. Constructive criticism will be encouraged throughout the progression of this study so the data presented be reflective of the scope of this study.

**Functional requirements**

The main program that will be used for the completion of this project are python, R and Tableau.

The link of the datasets that will be used to conduct this study is:

<https://www.kaggle.com/fedesoriano/stroke-prediction-dataset/version/1>

The variables that will be used to conduct this study are listed below:

**Independent Variable(s)**

**Variable**

ever\_married

□ Categorical: # of levels \_\_1\_\_\_ □ Continuous

**Variable**

work\_type

□ Categorical : # of levels \_1\_\_\_\_ □ Continuous

**Variable**

Residence\_type

□ Categorical: # of levels \_\_1\_\_\_ □ Continuous

Variable

Avg\_glucose\_level

□ Categorical: # of levels \_\_\_\_\_ □ Continuous

Variable

bmi

□ Categorical: # of levels \_\_\_\_\_ □ Continuous

Variable

smoking\_status

□ Categorical: # of levels \_\_1\_\_\_ □ Continuous

Variable

gender

□ Categorical: # of levels \_1\_\_\_\_ □ Continuous

Variable

age

□ Categorical: # of levels \_1\_\_\_\_ □ Continuous

Dependent Variable(s)

**Variable**

stroke

□ Categorical: # of levels \_\_1\_\_\_

**Variable**

heart\_disease

□ Categorical: # of levels \_1\_\_\_\_ □ Continuous

**Personnel requirements**

An estimated of 5 hour per day will be dedicated to conduct this study. Additionally, we have planned to meet with instructor Joseph Raetano every Monday at 8 PM (ET) to help us navigate through this journey of future data scientist.

**Delivery schedule**

Week 1: Import dataset into preferred software to begin data wrangling. Any unnecessary columns should be removed. Educate ourselves on breast cancer. Set up Github.

Week 2: Study the dataset and ask questions. What are some possible correlations? Is the data normally distributed? What are some predictive models we can make from it? Visualize the data to see if there is any interesting findings.

Week 3: Modeling/Optimization (Combined Stepwise - Forward and Backward Selection) and Machine Learning (Random Forest.)

Week 4: Review and validate findings from the previous week, and draw insights/conclusions.

Week 5: Compile findings into a Power Point slideshow. Go over it with their instructor and friend/family member to ensure that the presentation is clear and logical. Work on the style and layout of the presentation so it is delightful on the eyes.

Week 6: Make final touches to the Power Point presentation. Jane and Jessica should not attempt to come up with a brand-new analysis. There will not be enough time to verify their findings. They should practice presenting at least a couple times with the two of them, and at least once with their instructor.

**Other requirements**

All programs used should be free of charge.

**Assumptions**

To the best of our abilities, we will ensure that the software programs and platforms used are up to date.

**Limitations**

Scheduling and timing could be potential limitation to the completion of this project due to my availability. However, I will strive and make certain that I do communicate in a timely manner what I need to keep moving forward toward the completion of this study.

**Risks**

I am remaining optimistic that this project will get to completion in a timely manner. If anything, outside of my control arises, I will make certain to come up with an alternative solution so that this study does not get compromised.