Stroke Prediction Analysis Proposal

CDS-303: Scientific Data Mining

03-27-2023

Team Members:

Matthew Nguyen

Sulochan Panta

Yuta Sugiyama

Michael Tritle

Contents

[Introduction 3](#_heading=h.gjdgxs)

[Business problem and objective 3](#_heading=h.30j0zll)

[Analytic problem 3](#_heading=h.1fob9te)

[Goals and success criteria 3](#_heading=h.3znysh7)

[Resources available 3](#_heading=h.2et92p0)

[Requirements, assumptions, constraints 3](#_heading=h.tyjcwt)

[Conceptual model 3](#_heading=h.3dy6vkm)

[Data](#_heading=h.1t3h5sf) 4

[Data collection and cleaning](#_heading=h.4d34og8) 4

[Exploratory data analysis](#_heading=h.2s8eyo1) 4

[Feature selection / engineering](#_heading=h.17dp8vu) 4

[Handling and storing data](#_heading=h.3rdcrjn) 5

[Modeling](#_heading=h.lnxbz9) 5

[Machine learning methodology](#_heading=h.35nkun2) 6

[Tuning and testing plan](#_heading=h.1ksv4uv) 6

[Development environment and language](#_heading=h.rnrsb58kcwpr) 7

[Appendix A: Project Plan](#_heading=h.2jxsxqh) 8

[References 1](#_heading=h.3j2qqm3)2

# Introduction

Strokes are one of the leading causes of death in America, with more than 795,000 people every year having experienced one [1]. Noticing the signs of a stroke when it happens is one thing, but preventing it is another. Factors such as high blood pressure and weight may contribute heavily to ones’ chance of a stroke, therefore this analysis was conducted in order to find the most causal factor to a stroke rate and raise awareness on this subject.

## Business problem and objective

The business problem was stated as the following: “What lifestyle and genetic aspects contribute the most to stroke incidents?”

## Analytic problem

“If certain aspects were more apparent or were combined with other aspects of a subjects’ lifestyle and genetics, how would it affect the rate of having a stroke?”

## Goals and success criteria

The goal of this project was to find evident features that contribute to a stroke event in the hopes to raise awareness on the issue. Through supervised and unsupervised learning, the project is considered successful once the appropriate charts, models, and simulations all have the same conclusion and display the aforementioned goal.

## Resources available

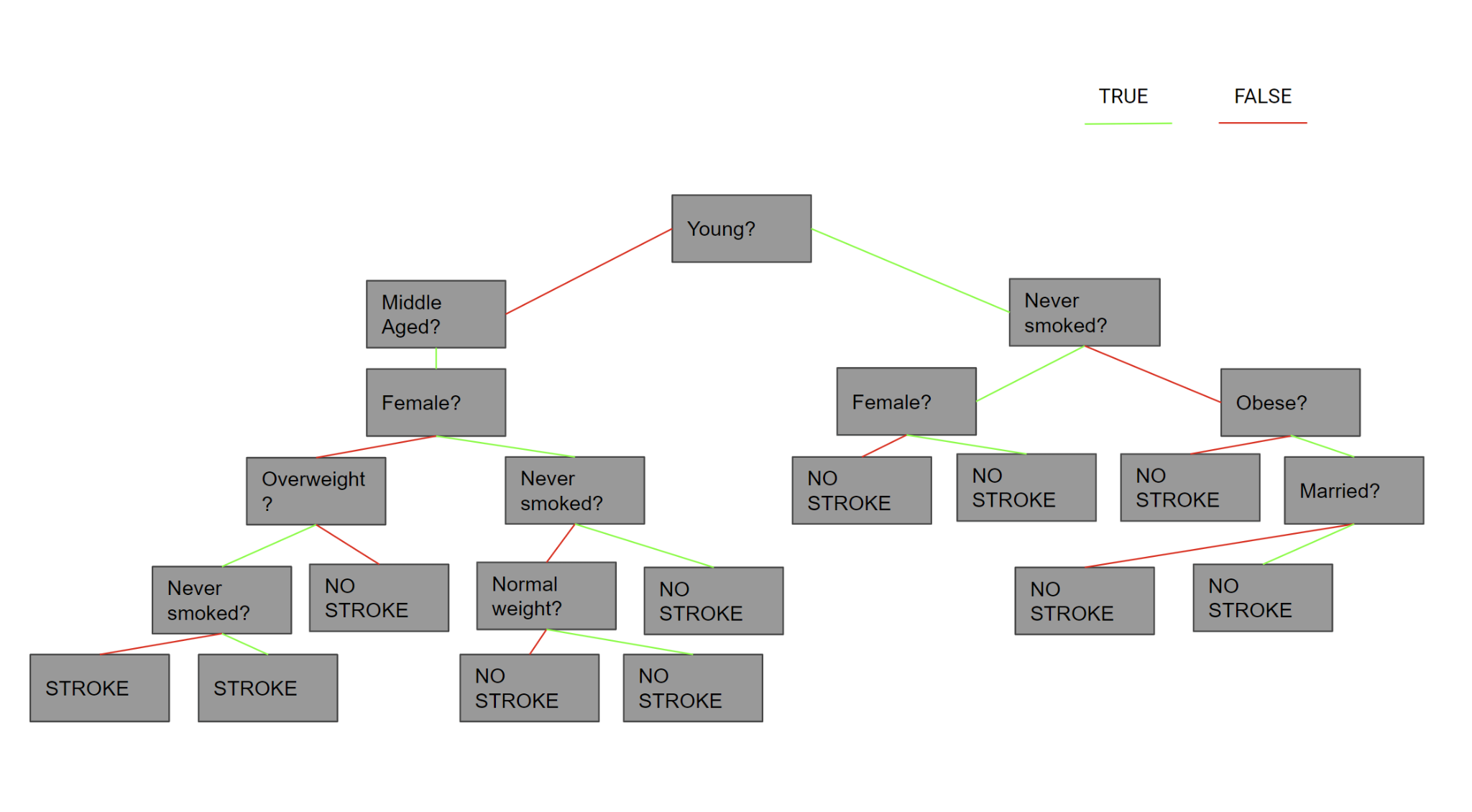
Main resource used in the project was a dataset on Kaggle named “Stroke Prediction Dataset,” last updated in 2021 [2].

## Requirements, assumptions, constraints

The first and most important requirement to the project is the Kaggle dataset, giving the team a foundation to build up on. One of the assumptions to be made without necessary research, is that high blood pressure and weight contribute heavily to stroke. However as the project progresses, this assumption could be proven to be true or not significant at all. Certain combination of attributes having more effect on the stroke event rate was another assumption made during the project. For example, any person who has a high BMI, smokes, and possesses an active history of heart disease would be assumed to be more susceptible to strokes than someone who has high rates on just two attributes. Male and female anatomy/lifestyles are historically different and having to compare BMIs and marriage status with each other may prove to be futile.

## Conceptual model

The decision tree diagram (Figure 1) displays a conceptual route of how the assessment can be made. As the tree goes further down, it would visually display what traits are responsible for stroke and whether the stacking of such traits increase the risk of stroke.



*Figure 1: Conceptual decision tree*

# Data

The dataset is used to predict whether a patient is more likely to get a stroke based on the input parameters: gender, age, heart disease, hypertension, smoking status, BMI, marital status, employment status, and average glucose level. The columns id, age, hypertension, heart\_disease, avg\_glucose\_level, bmi, are integer/float values, while the columns gender, ever\_married, work\_type, residence\_type are string values, which are encoded into various boolean columns. The columns heat\_disease and hypertension have integer values, but are transformed like the string values.

## Data collection and cleaning

The data, collected using kaggle, was first cleaned using python to fill in null values with the mean of the columns in the column, “bmi”. Then, by performing one-hot encoding, using mlxtend’s library, the data was encoded by breaking each unique category value out into its own column with boolean values.

## Exploratory data analysis

Exploratory data analysis was performed by creating a correlation matrix of all the classes (Figure 2). Additionally, two pie charts were created, with one representing the population percentage of each class related to individuals and their history of tobacco use (Figure 3); the second representing the population percentage of each individual’s occupation type (Figure 4).

## Feature selection / engineering

For feature selection, data columns such as employment status, average glucose level, bmi, and age, were reduced to their own class with boolean values. This also required balancing of data using a random sampler to make the code compatible with association rule and decision tree.

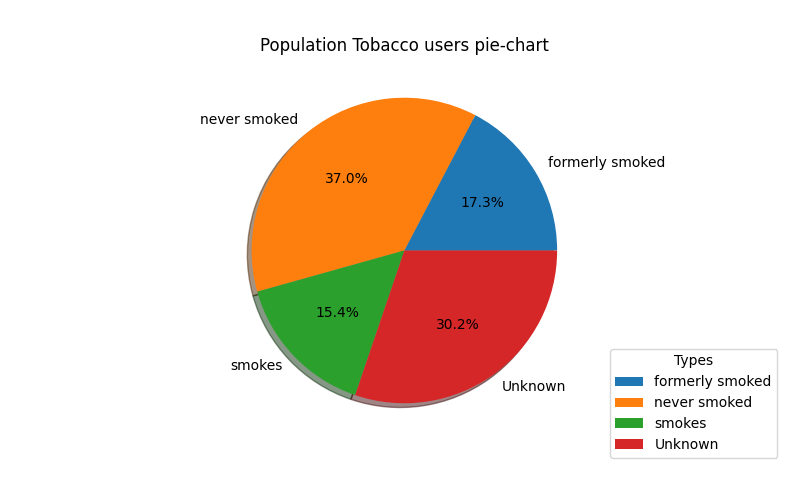
## Handling and storing data

The handling of the data was to balance the dataset between people who’ve had a stroke and a person that never had a stroke. Since the target classes were unbalanced, the non-stroke instances were randomly sampled until the target classes were at a 1:1 ratio.

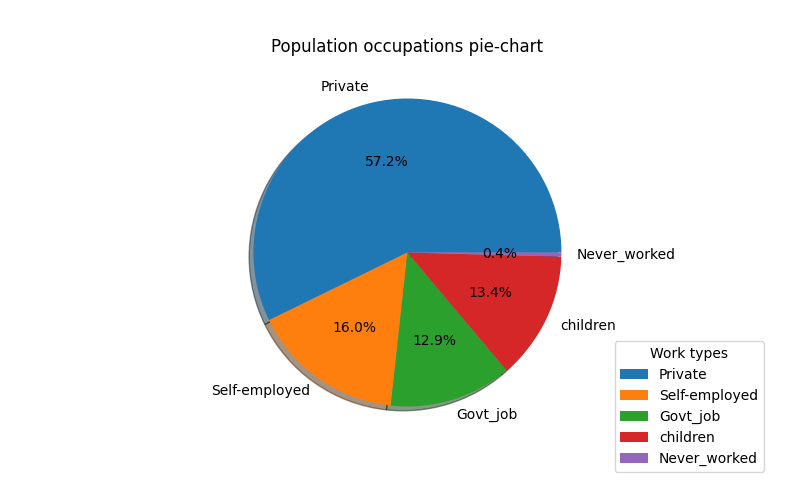
*Table 1: This is a table of the first 10 rows within the original dataset*

| id | gender | age | hypertension | heart\_disease | ever\_married | work\_type | Residence\_type | avg\_glucose\_level | bmi | smoking\_status | stroke |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 9046 | Male | 67 | 0 | 1 | Yes | Private | Urban | 228.69 | 36.6 | formerly smoked | 1 |
| 51676 | Female | 61 | 0 | 0 | Yes | Self-employed | Rural | 202.21 | N/A | never smoked | 1 |
| 31112 | Male | 80 | 0 | 1 | Yes | Private | Rural | 105.92 | 32.5 | never smoked | 1 |
| 60182 | Female | 49 | 0 | 0 | Yes | Private | Urban | 171.23 | 34.4 | smokes | 1 |
| 1665 | Female | 79 | 1 | 0 | Yes | Self-employed | Rural | 174.12 | 24 | never smoked | 1 |
| 56669 | Male | 81 | 0 | 0 | Yes | Private | Urban | 186.21 | 29 | formerly smoked | 1 |
| 53882 | Male | 74 | 1 | 1 | Yes | Private | Rural | 70.09 | 27.4 | never smoked | 1 |
| 10434 | Female | 69 | 0 | 0 | No | Private | Urban | 94.39 | 22.8 | never smoked | 1 |
| 27419 | Female | 59 | 0 | 0 | Yes | Private | Rural | 76.15 | N/A | Unknown | 1 |
| 60491 | Female | 78 | 0 | 0 | Yes | Private | Urban | 58.57 | 24.2 | Unknown | 1 |

*Figure 2: This is correlation matrix plot assist in predicting the evolution of the relationship between each class*

**

*Figure 3: This is a pie chart which depicts the scale between each smoke-related class in the dataset*



*Figure 4: This is a pie chart which depicts the scale between each occupation-related class in the dataset*

# Modeling

## Machine learning methodology

For this project we used Association rule mining and decision tree algorithms to have a predictive analysis and knowledge discovery from the provided dataset. This Project utilized the algorithms to find patterns and links between the variables in our data. To discover recurring patterns or connections between the variables in the data set, association rule mining was employed. It is used to determine the patient characteristics that are most likely to result in a stroke, such as age, gender, smoking status, and other illnesses. This can be helpful for early diagnosis and stroke risk reduction strategies.

## 

## 

## 

## 

## 

## 

## 

## 

## Tuning and testing plan

For the decision tree model, this project used a grid search on the following hyperparameters: “criterion”, “max\_depth”, “min\_samples\_split”, and “min\_samples\_leaf” to find the best hyperparameters.

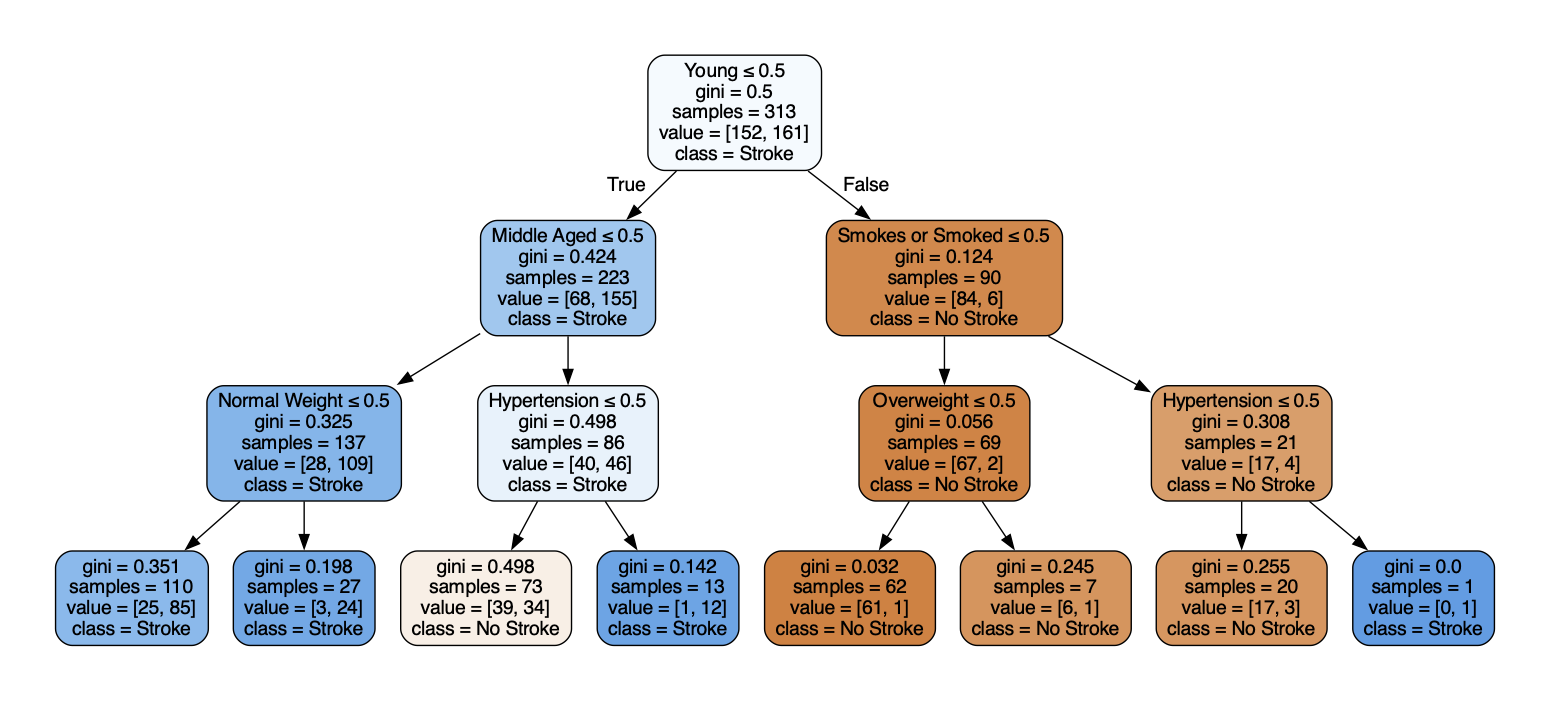
| “criterion” | “gini,” “entropy” |
| --- | --- |
| “max\_depth” | 2, 3, 4, 5, 6 |
| “min\_samples\_split” | 5, 10, 15, 20 |
| “min\_samples\_leaf” | 2, 3, 4, 5, 10, 15, 20, 25 |

*Table 2: Hyperparameter Values*

For the testing, we split the data where 75% was used for training the decision tree classifier. The remaining 25% of the data was subsequently used for testing the decision tree.

## Development Environment and Language

There were multiple Python libraries that were used for our analysis of the stroke data set. This project used pandas’ data frame to load the data set into our development environment. There were several methods that helped us get a “feeling” for what we are working with. Some of these methods were head(), describe(), info(), and isnull(); this was us to explore the data, and managing missing values. This project also uses the seaborn and matplotlib libraries to visualize the data and make it easier for it to show the results. For balancing the target class, This project used ibmlearn’s RandomUnderSampling library. Lastly, for modeling, Scikit learn’s Tree module were used for constructing and training a decision tree classifier (Figure 5).



*Figure 5: This is a decision tree which shows logical pathways for an individual to get a stroke.*

# Appendix A: Project Plan

| **Order** | **Task / subtask** | **Can it be started when?** | **Must it be finished when?** | **Expected duration (days)** | **Assigned (name)** | **Status (not started, in process, complete** |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | Choose Topic | 2/15 | 2/26 | 1 | Matthew Nguyen | Complete |
| 2 | Write executive summary report section | 2/15 | 2/26 | 1 | Matthew Nguyen | Complete |
| 3 | Business Problem |  |  |  |  |  |
| 4 | Write business problem report section | 2/26 | 2/26 | 2 | Sulochan Panta | Complete |
| 5 | Create one slide that has both the business and analytic problem | 2/26 | 2/26 | 1 | Sulochan Panta | Complete |
| 6 | Analytic Problem |  |  |  |  |  |
| 7 | Write analytic problem report section | 2/26 | 2/26 | 2 | Matthew Nguyen | Complete |
| 8 | Create one slide that has both the business and analytic problem | 2/26 | 2/26 | 1 | Matthew Nguyen | Complete |
| 9 | Project plan |  |  |  |  |  |
| 10 | Create project plan | 3/6 | 3/6 | 1 | Yuta Sugiyama | Complete |
| 11 | Write project plan report section | 3/6 | 3/6 | 3 | Yuta Sugiyama | Complete |
| 12 | Create one slide on project plan | 3/6 | 3/6 | 1 | Yuta Sugiyama | Complete |
| 13 | Data |  |  |  |  |  |
| 14 | Find data | 2/15 | 3/1 | 0 | Michael Tritle | Complete |
| 15 | Prepare data | 2/15 | 3/1 | 0 | Michael Tritle | Complete |
| 16 | Put data in a place where model can access it | 2/15 | 3/1 | 2 | Michael Tritle | Complete |
| 17 | Perform exploratory data analysis on data | 2/15 | 3/1 | 5 | Michael Tritle | Complete |
| 18 | Develop visualizations based on EDA | 2/15 | 3/1 | 5 | Michael Tritle | Complete |
| 19 | Write data section of report, include EDA visualizations and discussion | 2/20 | 3/1 | 3 | Michael Tritle | Incomplete |
| 20 | Create two slides on data | 2/20 | 3/1 | 1 | Michael Tritle | Incomplete |
| 21 | Methodology |  |  |  |  |  |
| 22 | Choose methodology | 3/6 |  | 1 | Matthew Nguyen | Complete |
| **Order** | **Task / subtask** | **Can it be started when?** | **Must it be finished when?** | **Expected duration (days)** | **Assigned (name)** | **Status (not started, in process, complete** |
| 23 | Write methodology section of report | 2/20 | 3/1 | 3 | Michael Tritle | Incomplete |
| 24 | Create one slide on methodology | 2/20 | 3/1 | 1 | Michael Tritle | Incomplete |
| 25 | Modeling (choice, development) |  |  |  |  |  |
| 26 | Determine appropriate model type | 3/27 | 4/3 | 1 | Yuta Sugiyama | Complete |
| 27 | Develop black box / conceptual model | 3/27 | 4/3 | 2 | Yuta Sugiyama | Complete |
| 28 | Develop model code based on black box model | 3/27 | 4/3 | 10 | Yuta Sugiyama | Complete |
| 29 | Write model type and justification report section | 3/27 | 4/3 | 3 | Yuta Sugiyama | Incomplete |
| 30 | Create one slide on modeling choice, also showing black box model | 3/27 | 4/3 | 1 | Yuta Sugiyama | Incomplete |
| 31 | Create one slide summarizing the modeling process | 3/27 | 4/3 | 1 | Yuta Sugiyama | Incomplete |
| 32 | Modeling (hyperparameter tuning) |  |  |  |  |  |
| 33 | Develop approach for tuning hyperparameters | 4/3 | 4/10 | 2 | Matthew | Incomplete |
| 34 | Tune hyperparameters of model using developed approach | 4/3 | 4/10 | 5 | Matthew | Incomplete |
| 35 | Write hyperparameter tuning report section of report | 4/3 | 4/10 | 3 | Matthew | Incomplete |
| 36 | Create one slide describing hyperparameter tuning process | 4/3 | 4/10 | 1 | Matthew | Incomplete |
| 37 | Modeling (testing) |  |  |  |  |  |
| 38 | Develop test plan for model | 4/3 | 4/10 | 5 | Sulochan Panta | Complete |
| 39 | Test model | 4/3 | 4/10 | 5 | Sulochan Panta | Complete |
| 40 | Write model testing section of report | 4/3 | 4/10 | 3 | Sulochan Panta | Incomplete |
| **Order** | **Task / subtask** | **Can be started when?** | **Must be finished when?** | **Expected duration (days)** | **Assigned (name)** | **Status (not started, in process, complete** |
| 41 | Create two slides: one on test process and one on results | 4/3 | 4/10 | 1 | Michael | Incomplete |
| 42 | Modeling (output) |  |  |  |  |  |
| 43 | Develop concept for a way for people to see the output of the model (user interface) | 4/10 | 4/17 | 3 | Michael | Complete |
| 44 | Develop front end code for visualization | 4/10 | 4/17 | 5 | Michael | Complete |
| 45 | Develop model output visualizations for report | 4/10 | 4/17 | 5 | Michael | Complete |
| 46 | Write modeling output report section | 4/10 | 4/17 | 3 | Michael | Complete |
| 47 | Create two slides showing model output visualizations | 4/10 | 4/17 | 1 | Michael | Incomplete |
| 48 | Modeling (assumptions and limitations) |  |  |  |  |  |
| 49 | Write report section on assumptions and limitations | 4/17 | 4/24 | 3 | Yuta | Complete |
| 50 | Create one slide on modeling assumptions and limitations | 4/17 | 4/24 | 1 | Yuta | Complete |
| 51 | Evaluation |  |  |  |  |  |
| 52 | Interpret model results in the context of the business problem. | 4/17 | 4/24 | 1 | Yuta | Incomplete |
| 53 | Write report section on evaluation | 4/17 | 4/24 | 3 | Yuta | Incomplete |
| 54 | Create one slide on evaluation | 4/17 | 4/24 | 1 | Yuta | Incomplete |
| 55 | Deployment and maintenance |  |  |  |  |  |
| 56 | Develop conceptual plan for deployment and maintenance | 4/24 | 5/1 | 4 | Matthew | Incomplete |
| 57 | Write section on deployment and maintenance | 4/24 | 5/1 | 3 | Matthew | Incomplete |
| **Order** | **Task / subtask** | **Can it be started when?** | **Must it be finished when?** | **Expected duration (days)** | **Assigned (name)** | **Status (not started, in process, complete** |
| 58 | Create one slide on deployment and maintenance plan | 4/24 | 5/1 | 1 | Matthew | Incomplete |
| 59 | Conclusion |  |  |  |  |  |
| 60 | Write conclusion of paper | 5/1 | 5/1 | 1 | Sulochan | Incomplete |
| 61 | Create one conclusion slide | 5/1 | 5/1 | 1 | Sulochan | Incomplete |
| 62 | Compile report sections and presentation slides, edit | 5/1 | 5/1 | 1 | Sulochan | Incomplete |
| 63 | Confirm completion of milestones at each step | 5/1 | 5/1 | 7 | Sulochan | Incomplete |

# References

[1]

CDC, “Stroke Facts | cdc.gov,” *Centers for Disease Control and Prevention*, Oct. 14, 2022. https://www.cdc.gov/stroke/facts.htm (accessed Mar. 26, 2023).

[2]

“Stroke Prediction Dataset.” https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset (accessed Mar. 24, 2023).