Term Project

**Intro**

I have chosen option A for this project, taking a dataset, and applying algorithms. The dataset I have chosen is a *“stroke prediction dataset”* uploaded by the user *fedesoriano*. My initial reasoning for this dataset is to determine if BMI relates to strokes. BMI doesn’t always correlate to health issues so I thought I would be interesting to see the effect of it, if any. But also, ultimately see if it is possible to train neural networks to predict someone having a stroke according to the different variables given in the dataset.

**Dataset**

This dataset was incomplete there were some holes in place of the data, thus would not work unless filled or removed from the dataset. Within the BMI column there were 201 missing entries, so as a solution I took the mean of the total BMIs and replace them in the missing entries. To note that this method does skew the data slightly though.

To understand the database a visual representation might be easier. I have made graphs for the most relevant data that pertains to mainly physical health.  
Within this dataset there are more Females than Males, with about a 58% being Female (There is also 1 entry of *Other*). And most of the ages being around 40-60. 90.25% has no hypertension, and 94.60% of the people don’t have heart disease either. Another important note is that 32.76% have smoked before/currently smoke and others never smoked or are unknown(the unknown this might skew the data little as well) . And lastly out of this group 4.87% of them has had a stroke.   
*(See figures 1-6 on figures page)*

**BMI and Average Glucose Levels**

Chart, scatter chart

Description automatically generatedAn important graph to consider in this data set is the BMI vs Average Glucose Level since these are the two variables that relate that can vary.

I have applied K-mean clustering to this scatter plot graph. Using k = 3. As we can see there are three groups. The data seems to show that even if you have a higher BMI that doesn't mean you have a higher average glucose level. This seems to be correct since BMI is not an accurate representation of a person's health overall*. (See next page for figure)*

Chart, scatter chart

Description automatically generated

Now we must relate these variables to strokes and see weather either of these variables attribute to strokes. To do this I have used a 3D scatter plot to add stroke to the z-axis. *(See figure on next page)*

The above k-clusters represent clusters of people and their Average Glucose Level vs their BMI and if they have had a stroke . As you can see on there are 3 different groups of people.

* The first group have a distribution of BMIs but their AGL(~50 - ~70) is relatively low. This can contribute to the number of strokes thus that group having the second most amount.
* The second group of people, they are the middle category, their AGLs(~70 - ~160) and BMI display they are the group with the least number of strokes.
* But the last group of people have a higher AGLs(~160 - ~260) and are the group with the most strokes.

**This data displays that although your BMI is higher doesn't mean you have a higher chance of a stroke. However, your AGL can represent a chance of a stroke.**

Chart

Description automatically generatedChart

Description automatically generated

We have provided multiple views of the same diagram to better show the centroids.

**Training the data**

Now using the data in the dataset, we will try to create neural network to try predicting strokes.   
So, within the program we will drop the *‘ID’* column since it is not needed, and we drop *‘stroke’* column since that is what we want to predict. We have also dropped *‘work\_type’, ’Residence\_type’, ’ever\_married’* since this is something that is not directly health related nor in our scope of our problem we want to solve. We will also be using 50% of the data train our neural network.   
We used Logistic Regression and Gradient Boosting since they preform the best in a smaller dataset.

Text

Description automatically generatedText

Description automatically generatedText, letter

Description automatically generatedThese were the results of the test:

It seems that the neutral network was fairly accurate in predicting stroke. Also reviewing the stoke percentage, it is very low to begin with, so this is more or less correct.

**Problems**

An issue I came across was how to represent the data, I took a lot of time looking through the matplotlib documentation to get things working correctly. I also installed seaborn to represent some of the graphs easily.

As mentioned before filling in the blank data points may contribute to some inaccuracy. Also, the unknown smoking status may also contribute to inaccuracy of the smokers affect on strokes, thus I didn’t go into depth with that portion.

**Summary**

According to our graph and k-mean cluster we can see that BMI does not relate to stroke chances however, as seen if the average glucose level is too high or too low the chance for a stroke increase. We also can see that our neural network on “gender”,” age”,” hypertension”,” heart disease”, “AGL” “BMI” and “smoking status” can successfully be predict strokes using Logistic regression and Gradient boosting.

# Figure Page

Figure 2

Figure 1

Chart, pie chart

Description automatically generatedChart, pie chart

Description automatically generatedChart, pie chart

Description automatically generatedChart, pie chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, histogram

Description automatically generated

*Figure 6*

Figure 4

Figure 5

Figure 3