Yee Shem:

(Slide 1) Hello, We are group 8 from lab group Z136 and today we will be presenting our SC1015 project on stroke prediction. My team includes Angelo, Brian and myself, Yee Shem.

(Slide 2)(click) According to the World Health Organisation, stroke is the second leading cause of death and the third leading cause of disability.

The Global stroke fact sheet released by the WHO also revealed that the lifetime risk of developing a stroke has increased by over 50% in 17 years
(click)This means that the risk of someone developing a stroke in their lifetime increased from (click) 1 in 6 to 1 in 4.

(Slide 3) From 1990 to 2019, there has been a (click) 70% increase in stroke incidence, (click) 43% increase in deaths due to stroke and (click) 102% increase in stroke prevalence.

Therefore, our group believes that the worldwide prevalence of stroke is an issue that is important to look into.

(Slide 4) Before we continue, let's learn more about stroke

(click)A stroke occurs when the blood supply to part of the brain is blocked or when blood vessels in the brain burst.

When this happens, the brain cells start to die and the longer the brain is deprived of oxygen, the more severe the damage can be.

(click)Such damages include memory loss, emotional problems and even paralysis.

Although stroke can happen to anyone at any age, there are certain factors that increase the chances of having a stroke.

(Slide 5) As the saying goes, Prevention is better than cure

It is necessary for one to know their own stroke risk factors so as to pick up healthier lifestyle habits as preventive steps for stroke.

And so, our problem definition is:
(click)Can we use one’s lifestyle to predict if they will have a stroke?

(Slide 6) For this project, we will be using a stroke prediction data set with (click) 11 clinical features for predicting stroke from Kaggle (click) The size of our data set is 5110 and these are the various numerical and categorical data used to predict strokes such as gender, age and glucose level.

(Slide 7)(click) These are our primary features before data cleaning
Now, let's move on to exploratory analysis and data preparation.

(Slide 8) (click)Firstly, we identified the data that can be dropped.
For example, the other gender has only one count, so it is statistically insignificant
(click)(click)The count of null BMI values is 201 which is only 4% of our entire dataset, thus we can remove it.
(click) So in total, we removed 202 rows of data.

(Slide 9) However, the (click) unknown smoking status makes up 1483 out of 5110 of our datasets, representing a significant percentage. It is also categorical data and can be treated as its own value which can be used for modelling later.
Thus, we have decided to keep it.

(Slide 10) Next, we categorised the continuous data, average glucose level into 3 categories.
(click)Normal, Prediabetes and Diabetes.

(click) Based on guidelines by the Centre for Disease Control and Prevention, a blood sugar level of 99 milligrams per deciliter and below is normal, 100 to 125 indicates prediabetes and above 126 indicates diabetes.

(click)Our data count is now 914 for diabetes, 1022 for prediabetes and 2972 for normal

This is to help us better understand what level of glucose is healthy and whether it contributes to the risk of stroke.

(Slide 11) Moving on to analytic visualisation

We first used univariate visualisation to analyse our continuous data, which is avg glucose level, BMI and age.
Using histogram and violin plot, we observed that
The majority of data on avg glucose level is within the normal range of 90-100

(Slide 12) The majority of data on BMI is in the borderline overweight range between 24-26

(Slide 13) And age is skewed very little, meaning that the data set has a fair distribution of all ages

(Slide 14) We also noticed a number of (click)outliers in our avg glucose level and age data from the boxplot

However, we have decided not to remove them as they are true outliers and they represent natural variations in the population.
Hence, they should be kept.

(Slide 15) Next, we did bivariate data analysis between the continuous data and stroke using a boxplot.
There are 3 things that we can observe here.
Firstly, avg glucose level has a positive correlation with stroke
Secondly, there is a slight correlation between BMI and stroke
And lastly, age has a positive correlation with stroke

(Slide 16) Finally, we conducted bivariate visualisation on our categorical data using a count plot
There are many things we can notice here.
For example,
The female gender has a higher stroke count than males. But that could be due to the greater number of females over males in the dataset.
The count plot also shows the unbalanced nature of our dataset.

(Slide 17) (click) To summarise our data cleaning and EDA
This is our updated attribute information with the inclusion of the new glucose_cat attribute.

(Slide 18) We found that our data is heavily imbalance as non-stroke data far outweighs stroke data
And thus, we are unable to do much feature selection for our modelling as we cannot tell which feature is helpful in predicting stroke.

Angelo:
(slide 19) Now I will be talking about how we prepared our models. **(CLICK)** First off we had to change the columns with categorical data in string form into categorical data in integer form for the functions to read it properly. **(CLICK)** You may have noticed we removed the column on residential_type. This is because from our initial data analysis using CountPlot, for each type of residence, the number of people that had a stroke was more or less the same, hence it can be concluded that residential type does not affect the chance of stroke.

(slide 20) We then did **(CLICK)** the train-test split on the new data frame with **(CLICK)** random_state set to true. This is because if random_state is false, it will pick the data from the top. We noticed that the data of patients who did not suffer a stroke was sorted at the top, hence bringing about the problem of uneven data. 42 is just a random number that we used which we will be consistently using for all train test splits in this project.

(slide 21) From our preliminary plotting of the binary classification tree, we noticed abnormal results. Upon further research, we realised it could be due to us having very few positive cases of stroke. **(CLICK)** In our data set the total number of patients who suffered a stroke was only 209 compared to the 4699 that didn't, which became 167 to 3759 after we did a train-test split. So we employed a technique called SMOTE which stands for Synthetic Minority Over-sampling technique. **(CLICK)** It is an oversampling technique where the synthetic samples are generated for the minority class, in this case, non-stroke data. This algorithm helps to overcome the overfitting problem posed by random oversampling. **(CLICK)** After using SMOTE, our train dataset is 7540 cases. Of these 7540 cases, 50% are patients with stroke and 50% without.

(slide 22) Since we have so much categorical data, a binary classification tree would be the ideal model. **(CLICK)** We chose to plot the binary classification tree with a depth of 20. The greater the depth, the greater the accuracy of the model. However, there is the problem of overfitting. It refers to the condition when the model completely fits the training data but fails to generalise the testing unseen data. To avoid this, we used a middle point where the depth is high enough to give good accuracy while not being too high that it overfits.

(slide 23) Given the test and train cases, **(CLICK)** we used the tree to predict if a patient will have a stroke for the test and train case individually. To check this data, we first checked the goodness of fit of the model. The train and test cases have 0.999 and 0.883 classification accuracy respectively, meaning that our model was well-trained. We also printed the sensitivity and confusion matrix for each case. Although our classification accuracy was high, it was not a good model for this problem since it had a low true positive rate, meaning it wasn't good at predicting stroke.

Brian:
(slide 24) After trying to use decision trees, we decided to try using a random forest model to try to improve our binary classification accuracy rate. **(CLICK)** A random forest model consists of multiple decision trees, where each decision tree searches for the next feature randomly to form a final prediction. When these decision trees are combined, randomly selecting features allows for a model that does not overfit while keeping classification rate accuracy high.

(slide 25) We were able to achieve a model accuracy of 0.925 and a True Positive Rate of 0 Although this model has better accuracy than the decision tree model, the True Positive Rate was still low and couldn’t successfully predict any strokes in the test cases.

(slide 26) We then decided to use hyperparameter tuning through randomized grid search to try to optimise our random forest model so that we could achieve a higher True Positive Rate. Randomized Hyperparameter selection was used as it was less time-consuming in searching for the optimal hyperparameters as compared to regular GridSearch. **(CLICK)** We found that these hyperparameters were best for our Random Tree model and trained it again with these hyperparameters. **(CLICK)**

(slide 27) Even after model hyper-tuning, our classification accuracy stayed around the same, although True Positive Rates improved slightly.

(slide 28) (CLICK) We conclude that our attempts have not been very successful….
Even though our classification accuracy was high for our Decision Tree Model and Random Forest hyptertuned models, it did not represent that we had actually predicted any actual strokes in a meaningful way **(CLICK)** as our True positive rates for prediction were very low. **(CLICK)** To answer our problem statement, we theorise that since the features in the dataset we used did not have any clear correlations for stroke likelihood and that since we cannot achieve a reliable True Positive rate in predicting stroke patients, that it might not be entirely possible to predict strokes with just one's lifestyle.

(slide 29) We conclude that **(CLICK)** extreme features like being heavier or having higher blood sugar might raise the chances of one having a stroke, but we must admit **(CLICK)** that it's still very difficult to predict strokes since they are so rare and can strike anyone! **(CLICK)** Additionally, strokes might also be due to preexisting conditions, which the dataset might not have accounted for!

(slide 30) (CLICK) (CLICK) Even though our project was not very successful in predicting stroke likelihood based on one's lifestyle, it should still serve as a reminder of what we should look out for if we want to prevent strokes.

(CLICK) That's the end of our presentation, thank you.