1.1. Background and problem

As time progressed, people in the modern cities tend to have a better quality of life, the choice of eating and entertainment became much more varied. However, the living habits of people became immoderate at the same time, people lack exercise, binge eating behavior, the daily routine was interrupted by work. Many diseases were derived from such bad living habits, such as heart disease, stroke, cancer ..... Among them, stroke is one of the disease killers in the world that have an obvious trend of the patient becoming younger on average. Stroke was no longer a label for elderly people, but also the younger generation. According to a study published in 2020, about 10% to 15% of strokes have occurred in the age of 15 to 50 of the patients. (Sullivan et al., 2020) Although the over rate of having a stroke is decreasing, there is a trend that the patients become younger, this is a serious problem that should be faced squarely. Stroke is the second leading cause of death in the world after heart disease, according to the World Health Organization (WHO), approximately 11% of people died of stroke since 2000. Stroke is also the third leading cause of disability in the world. This shows that stroke has many bad impacts on humans.

1.2. Definition

Stroke refers to a medical condition that causes a sudden loss of brain functions. (Fouche et al,2020) It has happened when the blood supply to the brain is cut off and causes the brain cells can’t receive enough nutrients and oxygen. As a result, brain cells died in the impacted area. The break-off of blood supply was due to two main reasons.

(1) Blockage of blood vessels to the brain. Atherosclerosis is the main cause of this. This condition is known as Ischemic strokes.
(2) Brain artery breaks suddenly, leading to excessive bleeding inside the brain. This condition is known as Hemorrhagic strokes.

BMI (body mass index) refers to a measurement of the body fat based on the weight and height of a person. With different BMI, people are classified into different stages. It should be cared if the BMI is too high as this indicates high body fatness, which may lead to several health problems.

1.3. Motivation

With the changes in living habits, stroke is no longer a disease that mainly happened in the elderly, but also a warning to young and middle-aged people. This should be cared about in society and people should pay more attention to it. There are many factors that lead to strokes, such as age, living habits, health conditions, and other slight factors. The purpose of this study is to find out the relevance of stroke with different features, to see whether such features have a strong relationship with stroke, this can help us be alert to this disease and have a deeper understanding of it.

1.4. Dataset & methodology

In this project, the dataset we chose is Stroke Prediction Dataset from kaggle, author fedesoriano. We used python as our main tools to do the data processing, visualization, model training.

2. Objectives

2.1. Seems the older people are, the higher risk they get of a stroke (Keun-Sik Hong et al.,2013). We want to explore the relevance between age and stroke.

2.2. Seems those with higher levels of glucose had an increased risk of stroke(Xuenan Peng et al.,2020). We want to explore the relevance between average glucose level and stroke.

2.3. Research found that the higher BMI, the higher risk they have stroke (Tobias Kurth et al.,2002). We want to explore the relevance between BMI (body mass index) and stroke.

2.4. To explore the relevance between gender and stroke

2.5. To explore the relevance between smoking status and stroke

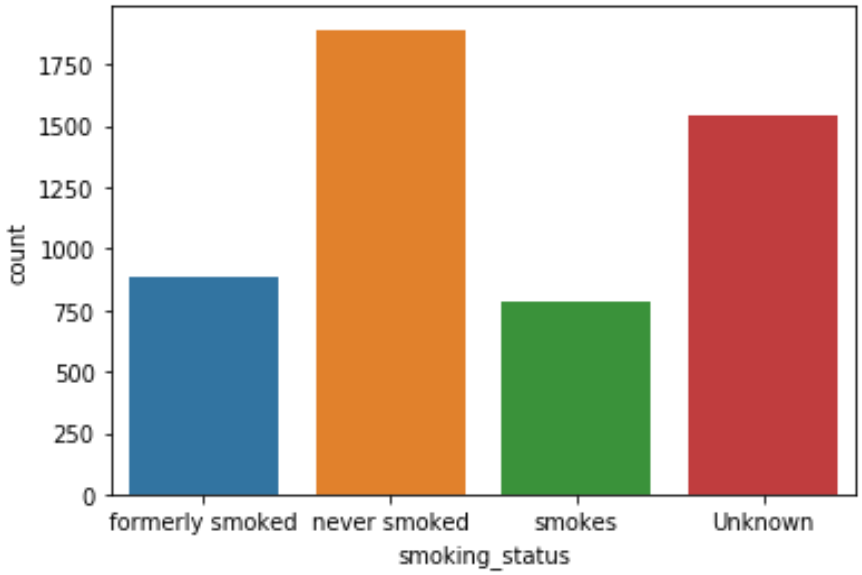
2.6. To explore the relevance between hypertension and stroke

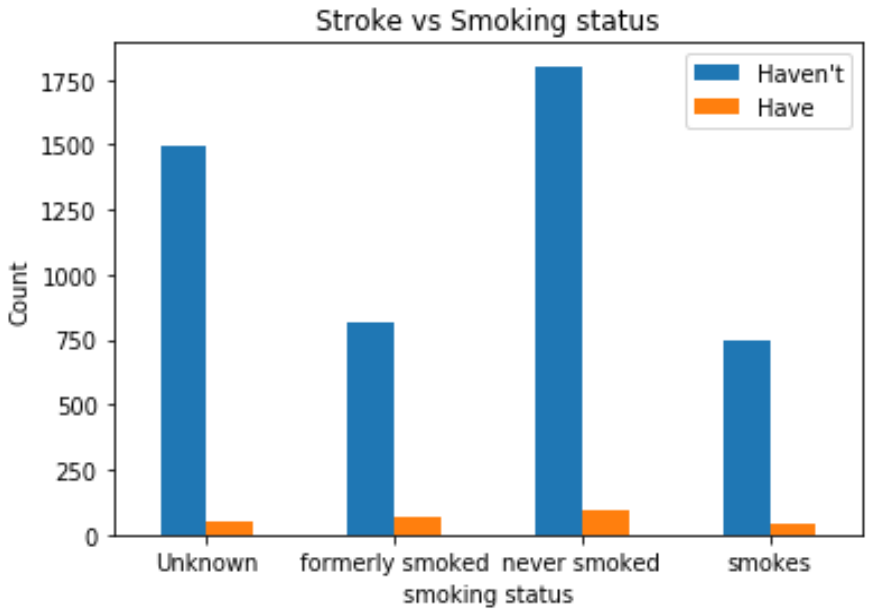
2.7. To explore the relevance between heart disease and stroke

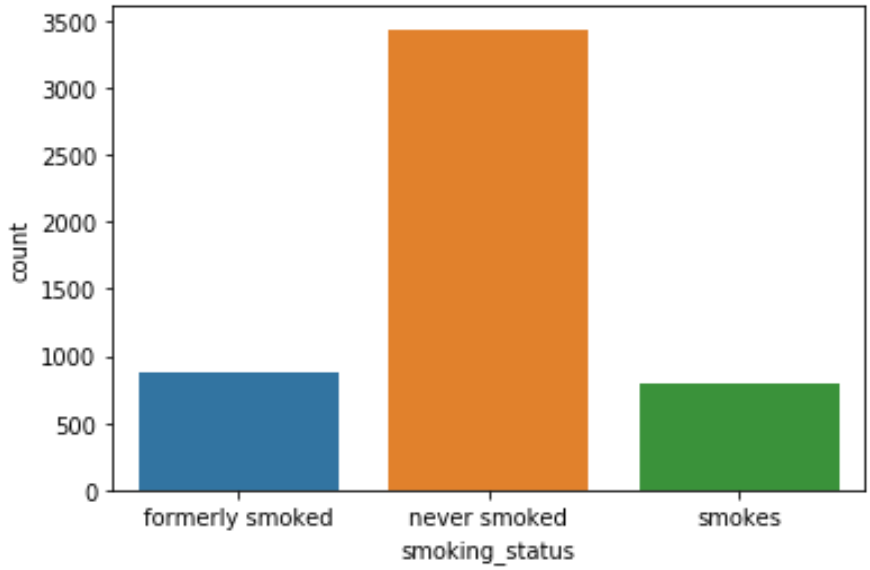
3. Data processing (Data cleaning)

We first check whether there are any missing data in the dataset. We found that 201 data about BMI is marked as N/A which are null values. The total dataset contains 5110 inputs. We replace them with the mean instead of deleting them as it is quite a great amount. Replacing by mean can also reduce the negative effects on the accuracy.

After that, we used visualization skills and found that there is a large amount of "unknown" for smoking status. We also treat unknown values as missing values. Then wevisualize the smoking status with a stroke graph to see whether unknown values took a vital role. The result is not. Therefore, we replace them with "never smoke" which is the mode so as to make sure the accuracy. The graph of unknown data and after cleaning result is shown below.





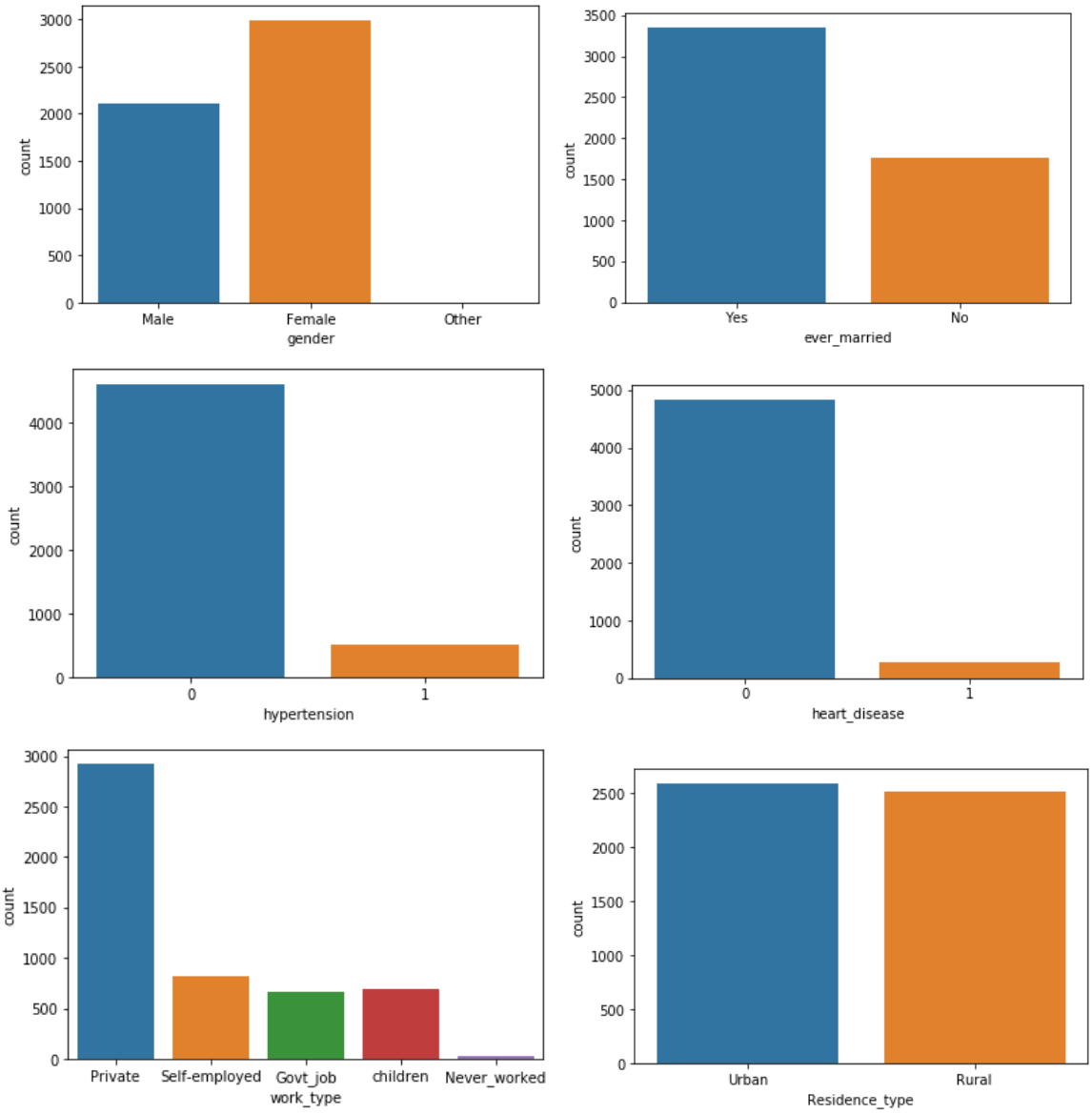


Lastly, we are going to prepare for the model training process. We delete "ID" from the database because it is not related to stroke and import scikit-learn library at python to do the label encoding. Because there are some inputs that are object type data which the model cannot be applied to if we do not transfer it into integer type.

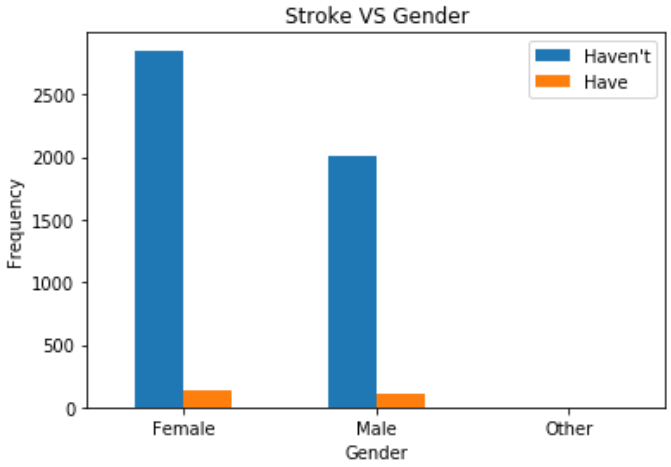
4. Data Visualization

In this part, we use graphs to show the distribution of people in different features that aim to find some relationship and trend.

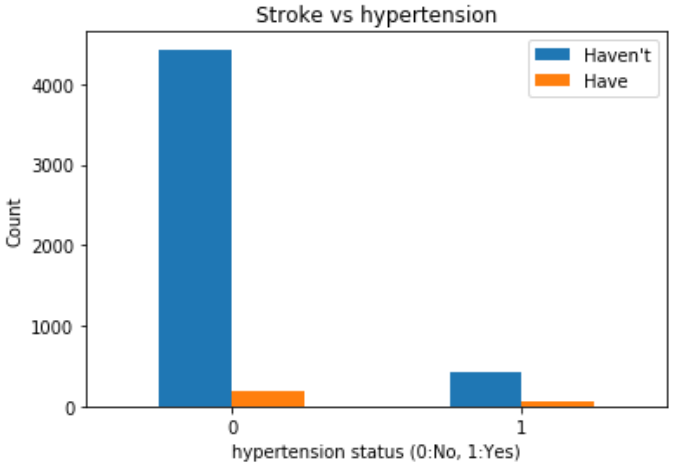
The bar charts below showing the distribution of gender, whether they were married, have heart disease or not,have hypertension or not, work type, residence type and smoking status with the x-axis is the count and the y-axis with the features.



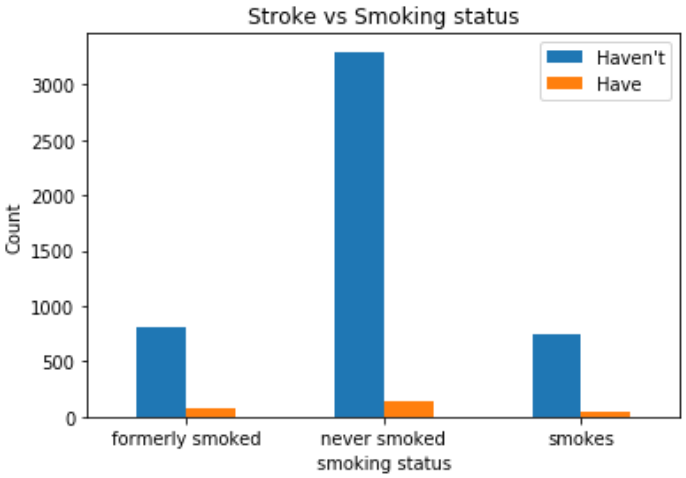
Next we also try to use graphs to show that how many people in each categorical attribute have strokes. The bar chart below shows the numbers of people who have or have not had strokes in each feature including gender, have hypertension or not, and smoking status with the x-axis is the count and the y-axis with the features.



In the graph of stroke vs gender, people who have a stroke are in small proportion when compared to those who do not have a stroke in both male and female.The amount of people having a stroke are similar in both male and female, which is around 100 to 200 people.

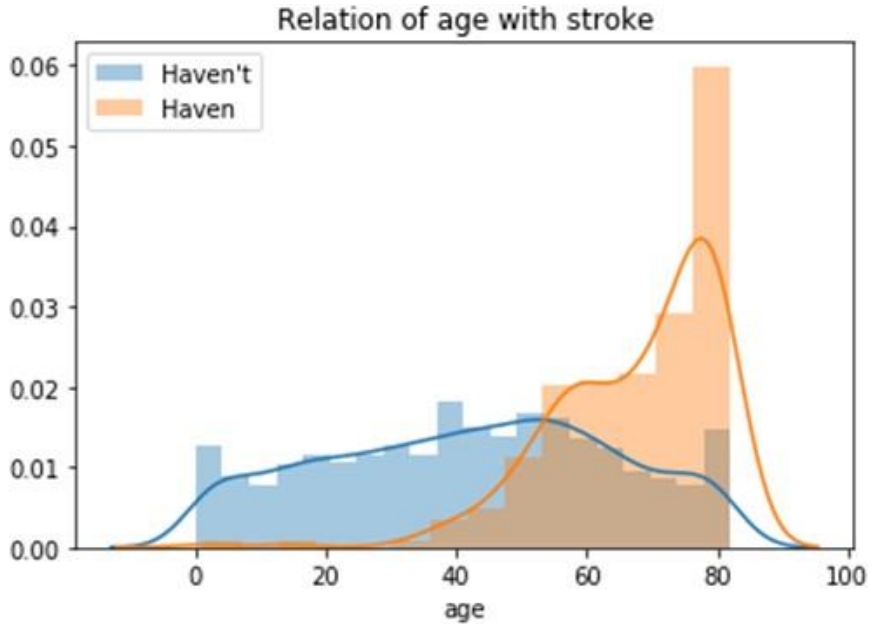


In the graph of stroke vs hypertension, there is a pattern that the amount of people who do not have a stroke are much more than those who have a stroke in both categories (having hypertension or not). However, in those people who do not have a stroke, most of them do not have hypertension while only a small proportion of them have hypertension.

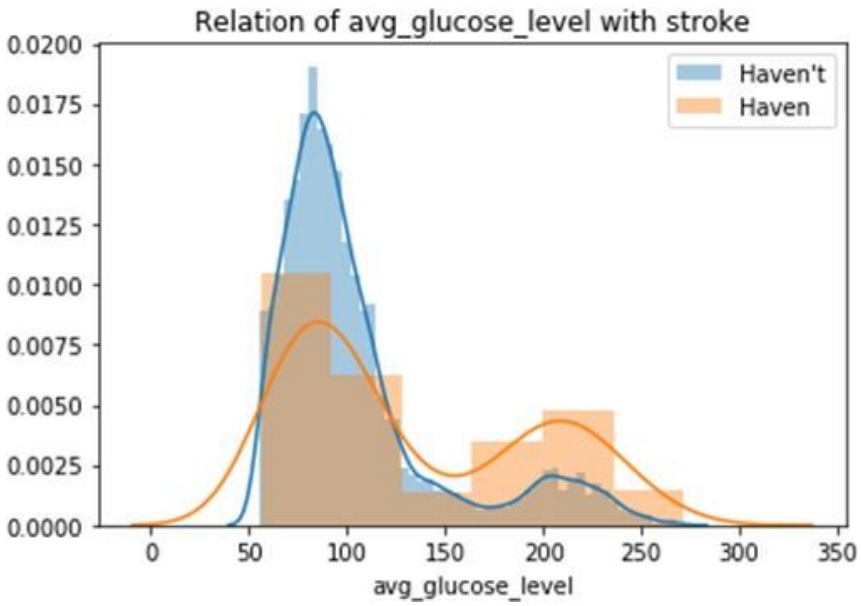


In the graph of stroke vs smoking status, people who do not have a stroke are mostly never smoked, which is around 3400. While people who formerly smoked and smoked share the similar amount in people who do not have a stroke, which is around 780 to 900. Among the three different smoking statuses, this graph tells that people who never smoked are less likely to have a stroke. But the dataset is kind of imbalanced, this conclusion does not give obvious convincing.

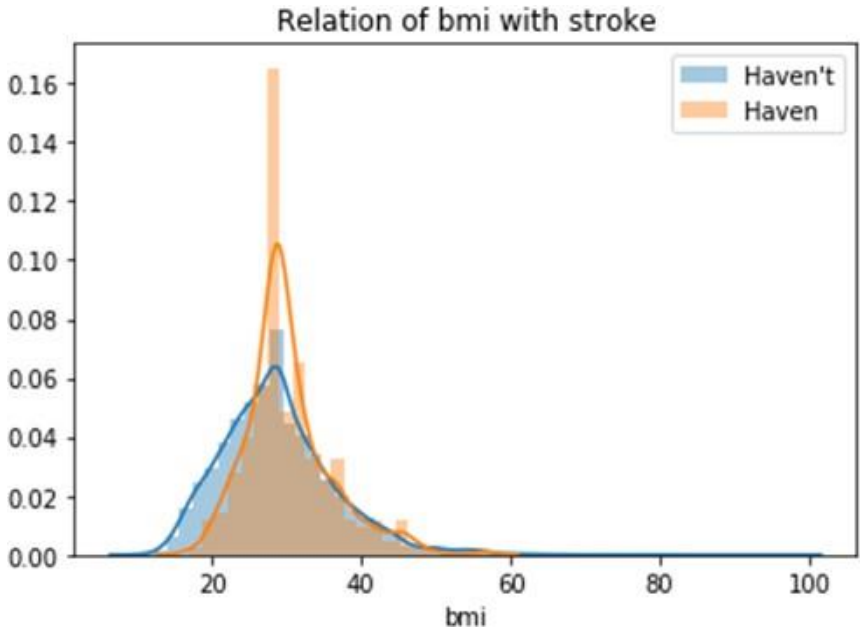
Next we also try to use graphs to show that how many people in each numerical attribute have stroke. The graphs below shows the numbers of people who have or have not had strokes in each feature including age, average glucose level and BMI.



From the above graph, we can see that most of the people who have strokes are above 40. When the age is increased, the more people have strokes. Also, when the age is increased, a larger proportion of people have stroke than the people who do not have stroke.



From the graphs above, we can see that for people who have the average glucose level that is in the range of 50 to 150, there are mainly a larger proportion of people who do not have stroke. However, for the people who have the average glucose level that is larger than 150, there is a larger proportion of people who have stroke. This shows that people who have a higher average glucose level have a larger chance to have stroke.



From the graphs above, we can see that for people who have a BMI below 30, there are mainly a larger proportion of people who do not have a stroke. However, for the people who have a BMI that is larger than 30, there is mainly a larger proportion of people who have stroke. This shows that people who have a higher BMI have a larger chance to have a stroke.

Here we have the visualization results:

- The data set contains 2115 male and 2994 female.
- The amount of people not having hypertension and and heart disease are in the majority and only a small proportion of people having those features.
- The work type of the respondents are mostly private.
- The two residence types (urban and rural) are nearly same in count.
- The people who have married are twice to those who have not ever married.
- When the age increases, there are more people who have stroke.
- The higher average glucose level and BMI, the higher chance for people to have a stroke.

5. Machine Learning Model

5.1. Model comparison

At the first stage, we did the feature selection and used the train test split function by an imported scikit-learn library. The dataset is randomly split into a train set, a test set followed by a ratio at 7:3, where the train set contains 3577 data, and the test set contains 1533 data.

```
x = en_data.drop('stroke', axis = 1)
y = en_data['stroke']
print('X Shape: ', x.shape)
print('Y Shape: ', y.shape)
# feature selection and see now how many rows and columns
# after drop stroke, only 10 types data
```

```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size= 0.3, random_state= 0)
# split data into 30% and 70&
```

```
print("Transactions: ")
print("x_train: ", x_train.shape)
print("y_train: ", y_train.shape)
print("x_test: ", x_test.shape)
print("y_test: ", y_test.shape)
# To see the shape of test dataset and train dataset
```

```
Transactions:
x_train:  (3577, 10)
y_train:  (3577,)
x_test:  (1533, 10)
y_test:  (1533,)
```

After we split the data, we tried to apply four different models including Random Forest Classifier, Decision Tree, k-nearest neighbors and Logistic Regression. We used the default valueto initialize all the hyperparameters of the model.

As we can see from Figure 1, there are 4 curves inside and each one represents one model performance. What we look for is AUC and EER. AUC stands for the area under the curve, and EER stands for equal error rate. Simply to say, the larger AUC and lower EER means better performance of the model. We found that the Random Forest Classifier gave the best performance over others, where the AUC is 0.82 and EER is 24.92%.

Therefore, we choose the Random Forest Classifier as our prediction model.

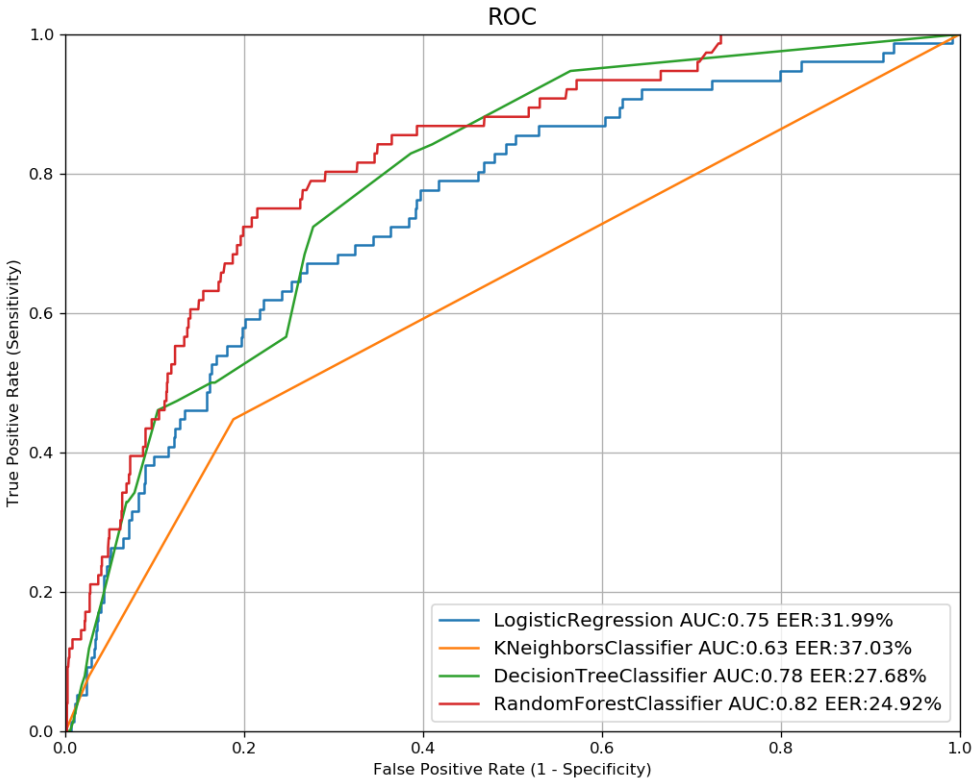


Figure 1. ROC of models

5.2. Selected Model

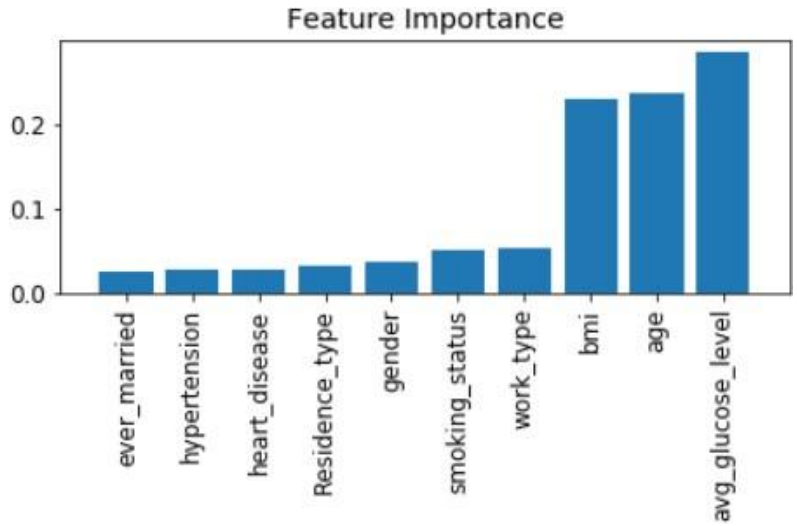
After we knew the Random Forest model provides the best prediction outcome, we would view the feature importances of the trained Random Forest Classifier. Figure 2 showed the feature importances of the trained model on the test set.

At the model comparison part, we have already splitted data into two parts, and did feature selection. To continue the process, we used standardization to give all the featuresthe same influence, then, in python, we import sklearn library and random forest classifier is utilized. Last, usd Random Forest feature importance as our result. We have plotted a picture to show that in the next slide.

As we can see from Figure 2, the top 3 features that have the highest relevance with strokes are Average glucose level, Age and BMI. It fulfills the reality that increasing age, high BMI will become much easier to get stroke. For other features, we can see there are relatively low importance.

1)	ever_married	0.023929
2)	hypertension	0.026358
3)	heart_disease	0.028379
4)	Residence_type	0.031781
5)	gender	0.036863
6)	smoking_status	0.050399
7)	work_type	0.053194
8)	bmi	0.228284
9)	age	0.235352
10)	avg_glucose_level	0.285462

Figure 2. Feature importance



6. Discussions

6.1. Result comparison

According to our objectives, we would like to explore the relationship between different features and stroke. We used data visualization techniques and a machine learning model method to do the prediction. And this part we will compare the visualization results with model prediction results.

As we found in visualization results, the elderly are tend to have strokes more than youngers. Especially after 60 years old, the chances to have a stroke is higher than younger. Compared with model results, it is clear to see that age is the second important feature that is relevant with stroke.

Same situation happened on average glucose level, it looks like a higher average glucose level, higher chance to have a stroke. And the model also showed us the same importance.

However, the visualization graph of BMI did not show it has an obvious phenomenon with stroke. But in our model prediction, it gave us a strong relationship with stroke, and this is fulfill the reality that higher bmi may increase the chance to have a stroke.

6.2. Limitation & improvement

In our prediction, there are some things that we should pay attention to, such as the size of the dataset and those missing and imbalanced values. It may be the reason that our model is not that precise.

In this dataset, every 5 out of 100 people are having strokes (Have: 4861, Haven't 249), which is highly imbalanced data. It is better to use oversampling to create more data next time, but due to our limited knowledge now, we do not know the exact concept behind, we hope we can do it after we learn.

There are still a lot of things we need to improve. Such as what we mentioned in the presentation that the missing value of smoking status also affects our results a lot. Afterwe looked deeper into the smoking data type and compared the smoking status with stroke statistics without data cleaning. It shows that unknown inputs do not really affect a lot. There are 1544 unknown which may affect our result a lot. We made a mistake while preparing the presentation slide. We should check carefully next time.

7. Conclusions

To conclude, after we did model and visualization comparison, it is clear that our model mostly matches the visualization results. To check objectives and references, we found that our prediction does indeed provide the proof. Like the older people are, the higher risk they get of a stroke and higher levels of glucose had an increased risk of stroke.

Our project is surely not perfect. We hope we can interpret balanced data like an oversampling dataset and apply more data science techniques next time.

8. References

Foucher, Gérard, & Faure, Sébastien. (2020). What is a stroke? *Actualités Pharmaceutiques, 59*(600), 57-60.

Keun-Sik Hong, Oh Young Bang, Dong-Wha Kang, Kyung-Ho Yu, Hee-Joon Bae, Jin Soo Lee, Ji Hoe Heo, Sun U Kwon, Chang Wan Oh, Byung-Chul Lee, Jong S Kim, Byung-Woo Yoon. (2013). Stroke statistics in Korea: part I. Epidemiology and risk factors: a report from the korean stroke society and clinical research center for stroke , *Journal of stroke 15 (1)* https://scholar.google.com.hk/sch²olar?q=stroke+and+age+and+BMI&hl=zh-TW&as_sdt=0&as_vis=1&oi=scholart#d=gs_qabs&u=%23p%3DMbjlmw_0uuwJ

Sullivan, K., Sullivan, K., Davidson, J. M., Rapaport, L., & Welch, A. (2020, June 21). *Think you're too young for a stroke? Think again: Everyday health*. https://www.everydayhealth.com/news/think-youre-too-young-stroke/.

Tobias Kurth, MD, MSc; J. Michael Gaziano, MD, MPH; Klaus Berger, MD, MPH; Carlos S. Kase, MD; Kathryn M. Rexrode, MD, MPH; Nancy R. Cook, ScD; Julie E. Buring, ScD; JoAnn E. Manson, MD, DrPH.(2002).*Body Mass Index and the Risk of Stroke in Men, Arch Intern Med*. ;162(22):2557-2562 https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/754810

World Health Organization. (2020, December 9). *The top 10 causes of death*. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death.

Xuenan Peng, Jinzhuo Ge, Congju Wang, Hongpeng Sun , Qinghua Ma, Yong Xu, and Yana Ma. (2020). *Longitudinal Average Glucose Levels and Variance and Risk of Stroke: A Chinese Cohort Study*, International Journal of Hypertension, vol. 2020, Article ID 8953058, 8 pages 2020https://doi.org/10.1155/2020/8953058

9. Appendix

Our Code for the project

Link:
https://portland-my.sharepoint.com/:u:/g/personal/haolonluo2-c_my_cityu_edu_hk/EXO6EWONGV5GtqZHpc_NcqIBhrJVNmv6eTi8y_PzqMS4pQ?e=IqbZuk

[1]:
```
import numpy as np
import pandas as pd
import os
for dirname, _, filenames in os.walk('D:\stroke\healthcare-dataset-stroke-data.csv'):
    for filename in filenames:
        print(os.path.join(dirname, filename))
```

[2]:
```
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

[3]:
```
df = pd.read_csv('D:\stroke\healthcare-dataset-stroke-data.csv')
df.head(10)
```

[3]:

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1
5	56669	Male	81.0	0	0	Yes	Private	Urban	186.21	29.0	formerly smoked	1
6	53882	Male	74.0	1	1	Yes	Private	Rural	70.09	27.4	never smoked	1
7	10434	Female	69.0	0	0	No	Private	Urban	94.39	22.8	never smoked	1
8	27419	Female	59.0	0	0	Yes	Private	Rural	76.15	NaN	Unknown	1
9	60491	Female	78.0	0	0	Yes	Private	Urban	58.57	24.2	Unknown	1

Stroke Prediction Dataset

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

```
df.head(10)
```

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1
5	56669	Male	81.0	0	0	Yes	Private	Urban	186.21	29.0	formerly smoked	1
6	53882	Male	74.0	1	1	Yes	Private	Rural	70.09	27.4	never smoked	1
7	10434	Female	69.0	0	0	No	Private	Urban	94.39	22.8	never smoked	1
8	27419	Female	59.0	0	0	Yes	Private	Rural	76.15	NaN	Unknown	1
9	60491	Female	78.0	0	0	Yes	Private	Urban	58.57	24.2	Unknown	1