M32084

Assignment 2:
Stroke Prediction Using Machine Learning

UP2067015

Table of Contents

Abstract

Introduction: Stroke is one of the popular killers in the world. Nowadays along with the breakthrough technologies and medical advancement, we obtain a deeper understanding of this disease. However, we are still improving our techniques on stroke prediction in order to help us take the right precaution at the right time.

Methods: In this article, we analysed a dataset that consists of potential factors of stroke by using IDE like Jupiter and exploratory data analysis. We also applied 3 different big data techniques, namely Random Forest, Logistic Regression, and Neural Networks to predict if a person suffered from stroke.

Result: The result suggests Logistic Regression is the better technique in this case. It also verifies the belief of some scholars that age and gender have fair influence on the risk for stroke.

Conclusion: The findings included in this article enhance our understanding of stroke and also help us realised that big data techniques can be very useful and practical in a way to help and save human lives.

1. Introduction

According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally. There are two ways that the brains blood supply can be interrupted and result in a stroke. The first is through a blocked artery called an ischemic stroke. This may be from a blood clot or a cholesterol plaque which blocks the vessels. The other type of stroke is from a burst artery that bleeds into the brain called a hemorrhagic stroke. When blood supply is interrupted through either of these two methods, some brain cells may die quickly yet others might last for hours. There's still some blood flowing around that area, as such after a stroke the extent of damage is present as what's called a cerebral infarction and which functions may be impacted can be highly varied. This is why after a stroke a series of tests and assessments are usually carried out.

Some scholars raised that risk of suffering from stroke does increase with age [2]. It's also more common in males [2]. Some scholars raised that risk of suffering from stroke does increase with smoking habit [3]. More studies can help us understand the relationship between stroke and age, gender, and smoking habit.

We will look into a stroke prediction dataset provided by *fedesoriano* [1] on Kaggle to see if we can successfully find out whether or not the aforementioned parameters and/or any other parameters in the dataset have impacts on the risk for stroke and by how much degree they affect it, by using 3 different machine learning techniques: Random Forest, Logistic Regression and Neural Networks.

2. Dataset

We have a multivariate dataset that contains of 12 features, 5110 individual entries in total. The 12 features include 3 numerical: 'age', 'avg glucose level' and 'bmi'; the remained are categorical: 'id', 'gender', 'hypertension', 'heart disease', 'ever married', 'work type', 'residence type', 'smoking status', and 'stroke'. The original table looks like this (top 5 rows):

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.	formerly smoked	1
51676	Female	61.	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
31112	Male	80.	0	1	Yes	Private	Rural	105.92	32.	never smoked	1
60182	Female	49.	0	0	Yes	Private	Urban	171.23	34.	smokes	1
1665	Female	79.	1	0	Yes	Self-employed	Rural	174.12	24.	never smoked	1

Since 'id' does not have any implication to other features besides as a unique number that helps identifying the entries, we decided to drop that column. Now our table looks like this (top 5 rows):

gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke

Male	67.(	0	1	Yes	Private	Urban	228.69	36.(	formerly smoked	1
Female	61.(	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
Male	80.(	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
Female	49.(	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
Female	79.(	1	0	Yes	Self-employed	Rural	174.12	24.(	never smoked	1

After that, we investigated if there is any missing values in our dataset, since they can affect the reliability of our machines learning models. We found 201 missing values under column 'bmi':

```
gender                0
age                   0
hypertension          0
heart_disease         0
ever_married          0
work_type             0
Residence_type        0
avg_glucose_level     0
bmi                 201
smoking_status        0
stroke                0
dtype: int64
```

Instead of imputing by the mean or median, we use a decision tree model to give values to those who are missing based on "age" and "gender" from the other samples. We now have a set of clean data.

```
gender              0
age                 0
hypertension        0
heart_disease       0
ever_married        0
work_type           0
Residence_type      0
avg_glucose_level   0
bmi                 0
smoking_status      0
stroke              0
dtype: int64
```

3. Exploratory Data Analysis

This is a binary classification, result in either "stroke" (stroke = 1) or "no stroke" (stroke = 0). We first explored the target feature, "stroke". Figure 3a shows that stroke distribution is highly imbalanced in our dataset.

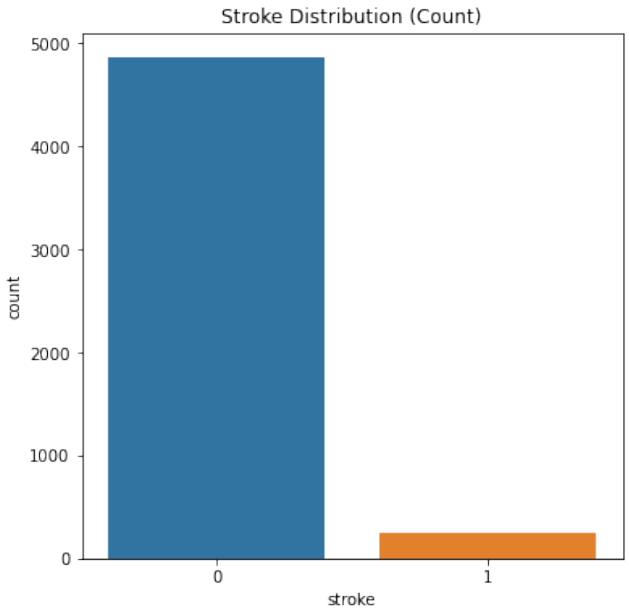

Figure 3a. Stroke Distribution (Count)

We changed to look at the percentage (Figure 3b) and found that less than 5% of our dataset represents the present of our target feature "stroke" (stroke = 1).

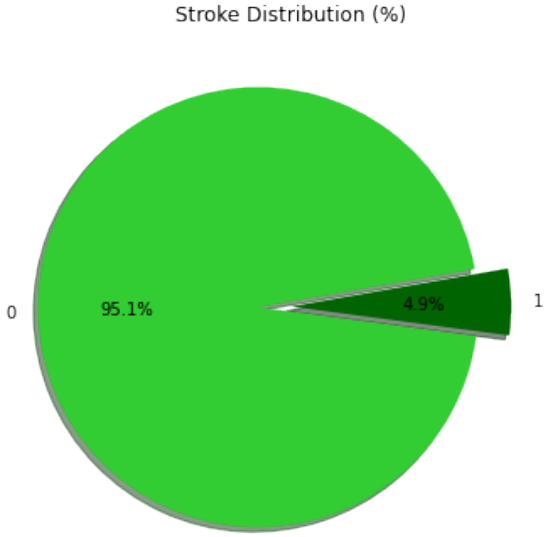

Figure 3b. Stroke Distribution (%)

That is nearly every 1 out of 20 people affected by stroke (Figure 3c).
Let's further look into how other features in relationship to stroke.

Figure 3c. Ratio of people affected by stroke

Figure 3d shows the relationship between 'stroke' and 'age'. We can see that the data with absence of stroke is fairly balanced, while the data with present of stroke seems left-skewed. It suggests that the risk for stroke likely increases with the increase of age. Therefore, "age" remains a parameter that worth further investigation.

Figure 3d. Stroke by Age

This figure also suggests that there are some cases of stroke happened under age of 20. Let's take a look of the data:

	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
162	Female	1.32	0	0	No	children	Urban	70.37	18.52	Unknown	1
245	Female	14.0	0	0	No	children	Rural	57.93	30.90	Unknown	1

The table indicates that a child of less than 2 year-old and a teenager of 14 year-old suffered stroke. We need to bear in mind that the reasons behind can be a wide guess. They are likely outliers, however.

Next we have a figure of stroke by average glucose level (Figure 3e). Of the people never experienced stroke, it happened to have a larger distribution of lower average glucose level (the first peak: 0-150)

than the higher (the second peak: 151-300), compared to that of the people experienced stroke. This suggests that people who experienced stroke were more likely to have higher average of glucose.

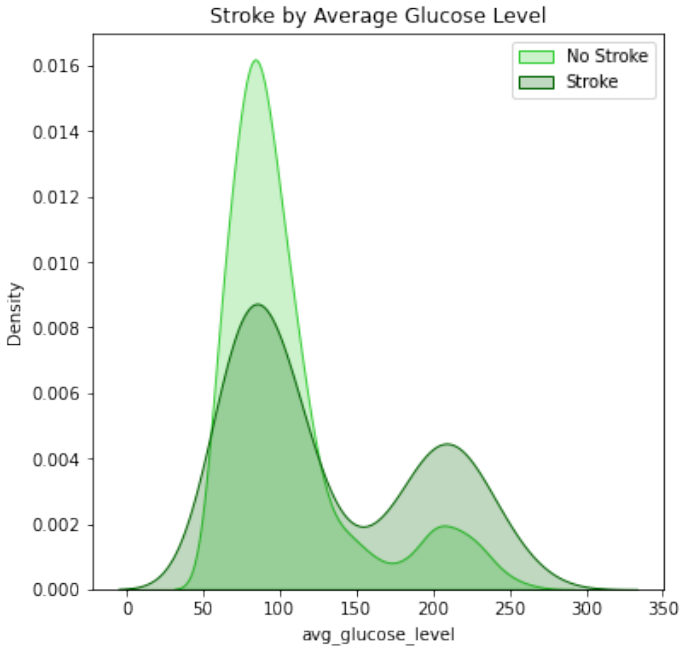

Figure 3e. Stroke by average glucose level

Then we look into the last continuous feature, the 'bmi'. From figure 3f, we can see that the shapes of distribution of both "No stroke" and "Stroke" are fairly similar. However, the distribution of "Stroke" is more centralised to 30 BMI.

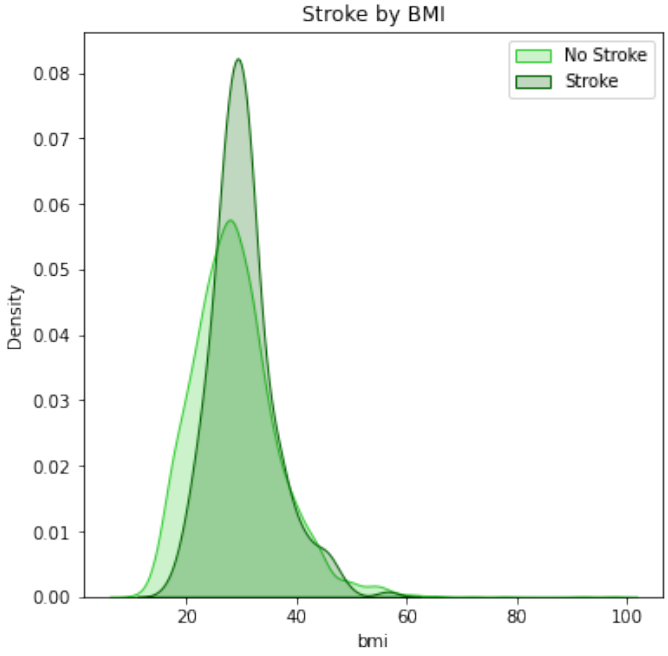

Figure 3f. Stroke by BMI

Now let's look into the categorical features.

First we have "gender", one of the features we mentioned in the introduction section that is deemed by scholars as one of the impacts to stroke. The histogram (Figure 3g) shows obvious differences of ratio — The numbers of reported stroke are nearly identical in both genders, while the total number of reported female in our dataset is obviously larger. This suggests that male is more likely to experience stroke than female.

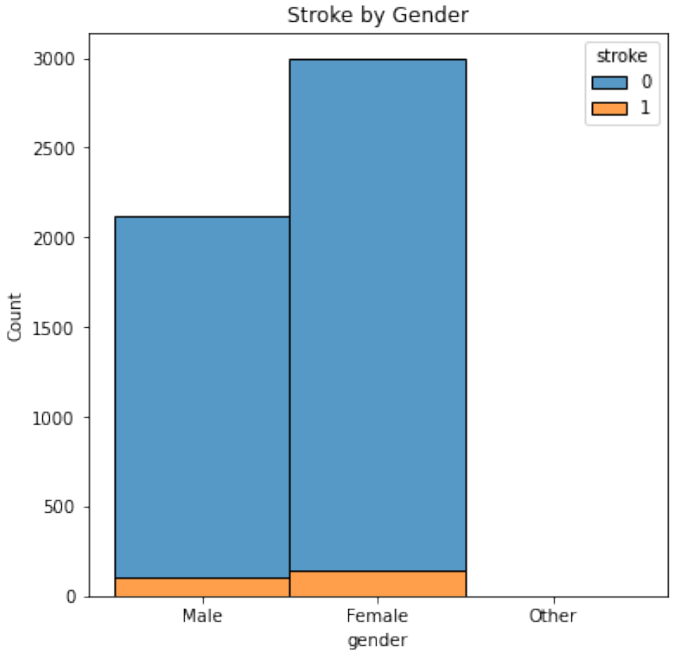

Figure 3g. Stroke by Gender

We also noticed that there is a category named "Other" of barely any visible case. We looked further into our dataset:

```
gender
Female    2994
Male      2115
Other        1
dtype: int64
```

Since there is only 1 case of "Other", we decided to drop it to keep our dataset simple.

We are also interested to compare the experience of stroke in both genders with the increase of age. By comparing the orange bars ('stroke' = 1) of the histograms below (Figure 3h), one thing we notice is that even female has earlier cases of stroke positive recorded, male tend to step into a higher risk for stroke in their late 50/early 60, while it is relatively late for female (Their median seems to be around age 70).

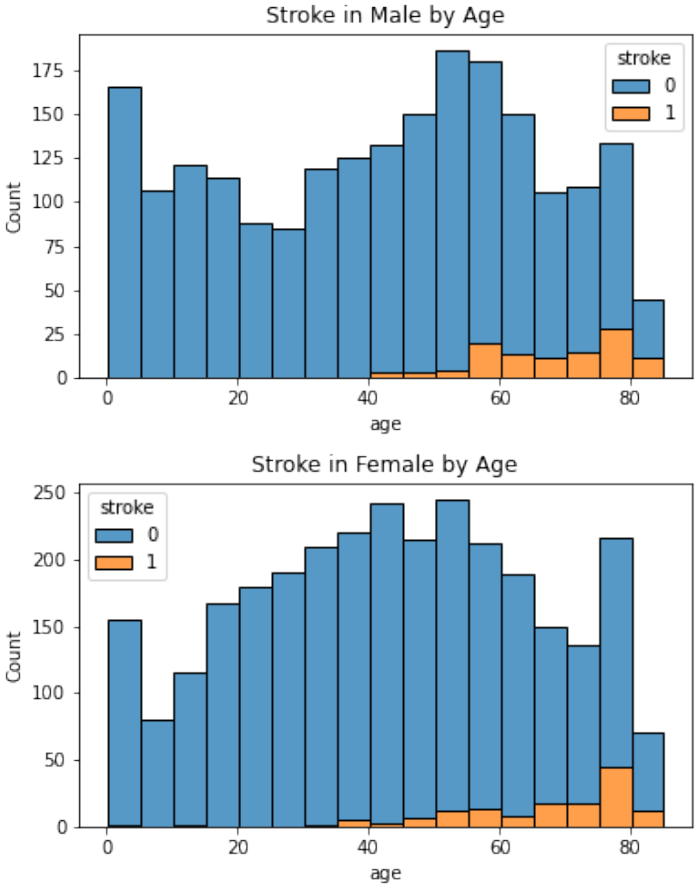

Figure 3h. Stroke by Gender by Age

We then examine the relationship of "stroke" and "hypertension".
13.3% of the people who have hypertension reported for stroke, while that number only accounts for 4% of the people who do not have hypertension (Figure 3i). It seems that the likelihood of having stroke is much higher in those who have hypertension than those who do not.

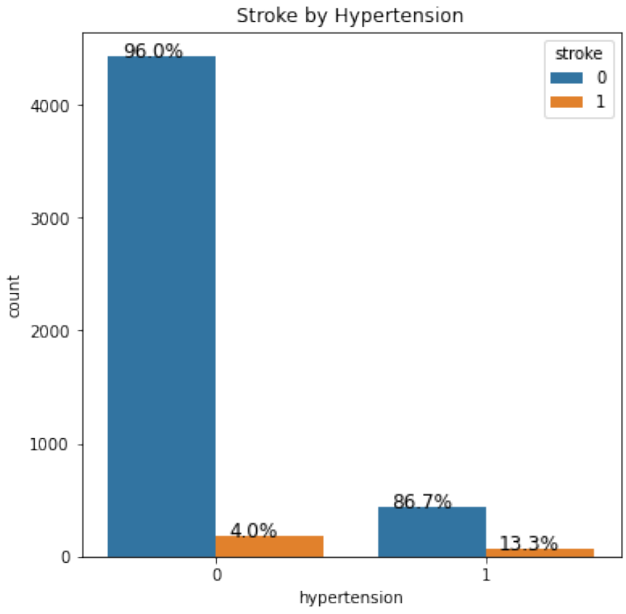

Figure 3i. Stroke by Hypertension

Next we examine the relationship of "stroke" and "heart disease".
17% of the people who have heart disease reported for stroke, while that number only accounts for 4.2% of the people who do not have heart disease (Figure 3j). It seems that the likelihood of having stroke is much higher in those who have heart disease than those who do not.

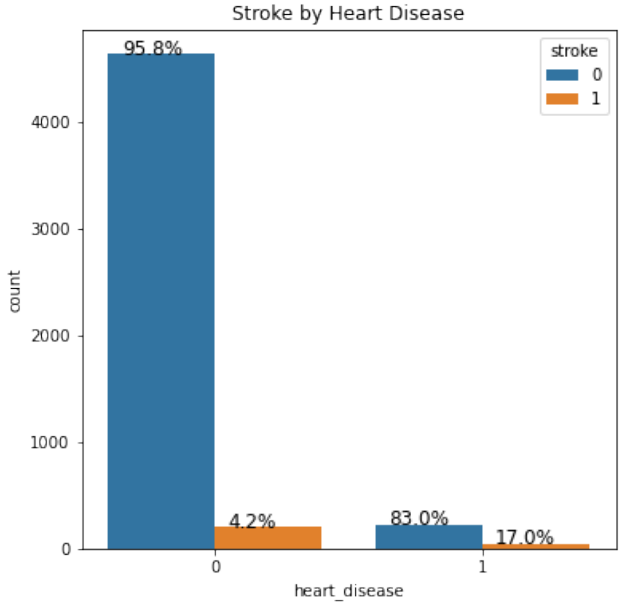

Figure 3j. Stroke by Heart Disease

After that, we examine the relationship of "stroke" and "ever married".
6.6% of the people who have ever married reported for stroke, while that number only accounts for 1.7% of the people who have never married (Figure 3k).

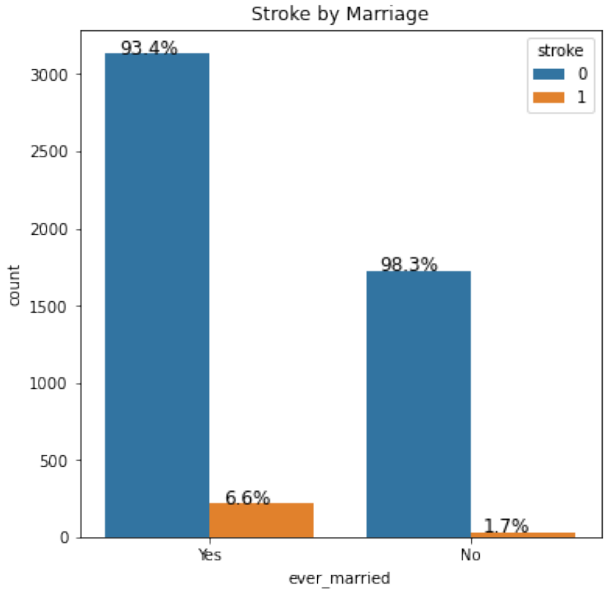

Figure 3k. Stroke by Marriage

It seems that the likelihood of having stroke is somewhat higher in those who have ever married than those who have not. Family and economic stress might have played an important role on it.

Next we examine the relationship of "stroke" and "work type".
"Private", "self-employed" and "govt job" have similar ratios of people reported for stroke while that number only accounts for less than 1% of the people who are "children" and "never worked" respectively (Figure 3l).

We do not have sufficient information to say that this is because (1) those reported "never worked"are in their young ages and thus "age" plays a role in that along with "children" or (2) people reported working ("Private", "self-employed" and "govt job") gain stress from their work or (3) any other reasons.

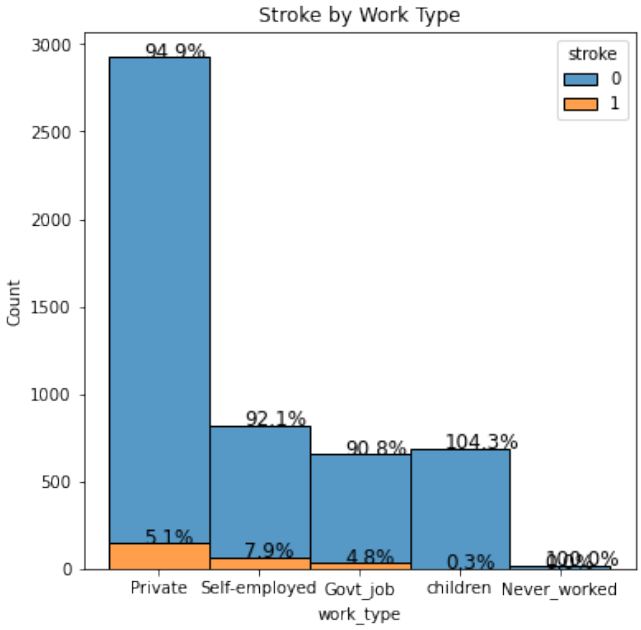

Figure 3l. Stroke by Work Type

Next we examine the relationship of "stroke" and "residence type".
People who lived in "urban" and "rural" seems to have similar level of occurrence ratio for stroke (Figure 3m).

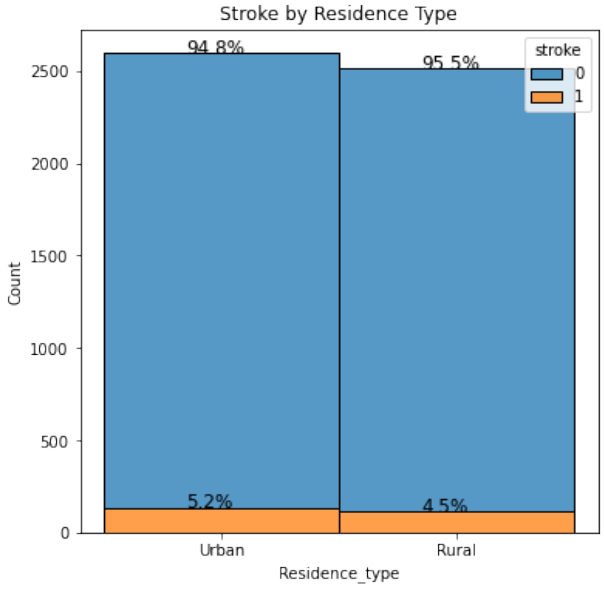

Figure 3m. Stroke by Residence Type

The last parameter we examine against "stroke" is "smoking status". This is one of the aforementioned traits that scholars believe to be impactful to the risk for stroke.

Surprisingly, among people reported "formerly smoked", "never smoked" and "smoke", "never smoked" slightly surpasses "formerly smoked" and "smoke" by 2.1% and 1% respectively (Figure 3n). This goes against what the scholars believe. However, the relationship between smoking and stroke remains in doubt since the people reported "unknown" accounts for the most amount (6%).

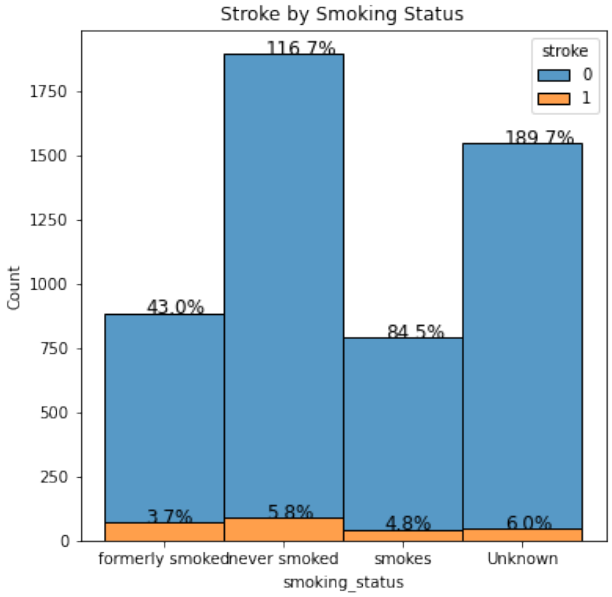

Figure 3n. Stroke by Smoking Status

Now, after analysing the relationship between stroke and those features, let's take a look at the correlation heat map (Figure 3o):

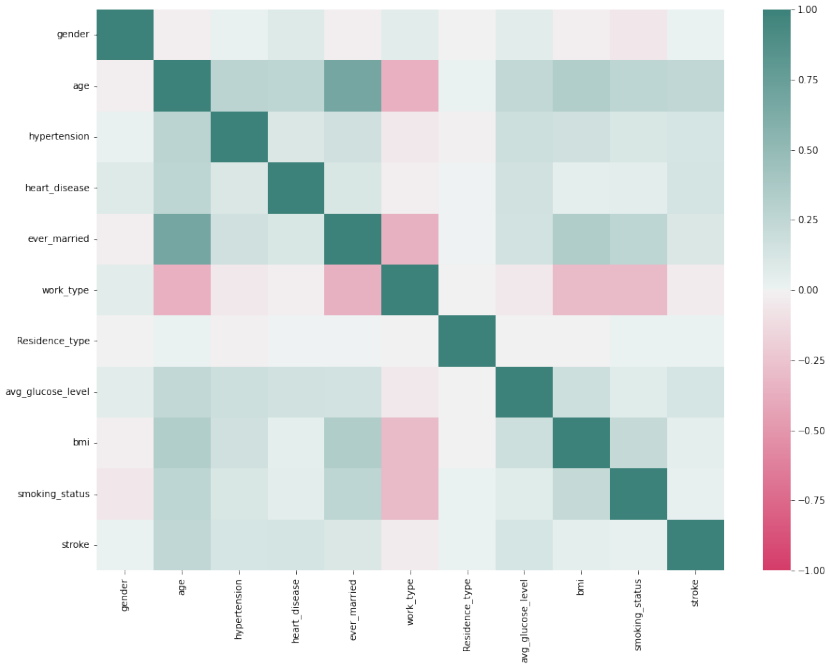

Figure 3o. Correlation (all features)

"Age" is the most correlated parameter to "stroke", followed by "hypertension", "heart disease", "ever married" and "glucose level" (almost on the same degree). Surprisingly, "gender", "work type", "residence type", "bmi" and "smoking status" show little to no correlation to stroke.

4. Preprocessing

Since our machine learning techniques allow only numerical features for calculation, thus we need to transform our categorical features into that. We decided to do so by applying encoding on the categorical features. In addition, our numerical features are on different scales. To avoid dominance in magnitude by one or more features, we applied scaling on the numerical features. The result is as follow (top 5 rows):

	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	1	1.051	0	1	1	2	1	2.706450	0.993	1	1
1	0	0.785	0	0	1	3	0	2.121652	0.124	2	1
2	1	1.626	0	1	1	2	0	-0.004867	0.463	2	1
3	0	0.255	0	0	1	2	1	1.437473	0.708	3	1
4	0	1.581	1	0	1	3	0	1.501297	-0.63	2	1
5	1	1.670	0	0	1	2	1	1.768299	0.010	1	1

Secondly, we separated the dataset into non-target ("X") and target sets ("y," contains only "stroke"). Then we split the sets into training and testing sets by a 70/30 ratio respectively. Now we have a total of 4 sets of data ready for use.

But before going into the next step, as we mentioned during preprocessing, our dataset is heavily imbalanced. We decided to use the package SMOTE only on our training sets to increase the amount of data points that are stroke positive in order to make our dataset balanced before being plugged for model training.

Now we are good to go.

5. Train and Test Models

We are going to use Random Forest, Logistic Regression, and Neural Networks for this task. In addition, we will use 10-fold cross validation method in order to obtain a better representative of model performance compared to less-fold validation.

The scores of 10-fold cross validation are shown in figure 5a. Both models performed fairly consistent most of the time. It is easy to find that Random Forest is the better performer here.

Figure 5a. 10–Fold Cross Validation Scores

5.1. Random Forest

Let's deal with the Random Forest model first.
From the above 10-fold cross validation, we obtained the mean F1 score as follow:

```
Random Forest Mean f1 score: 0.1496881496881497
```

Next step, we are going to plug the data created by the package SMOTE into the trained models to see which models perform better with unseen data.

5.1.1. Preliminary Result and Evaluation

Let's have a preliminary evaluation for the model:

	precision	recall	f1-score	support
0	0.96	0.92	0.94	3404
1	0.12	0.21	0.15	173
accuracy			0.89	3577
macro avg	0.54	0.56	0.54	3577

```
weighted avg       0.92      0.89      0.90      3577

Accuracy Score:  0.8856583729382164
F1 Score:  0.1496881496881497
```

The model gives us a high accuracy of 0.89 and a disappointing recall. It also records a significant drop in F1 score. Let's give it a parameter tuning to obtain a better performance.

5.1.2. Parameter Tuning

We used the package GridSearchCV to execute an exhaustive search over parameter values.
We searched the number of estimators among 60, 100, 120 and 180 and got the best outcome at 100.
We searched maximum features among 2, 3, 7, 9 and 15 and got the best outcome at 2.
We searched both with and without bootstrap and got the best outcome with bootstrap.

5.1.3. Final Result and Evaluation

Below is the final result we obtained:

```
              precision    recall  f1-score   support

           0       0.96      0.92      0.94      3404
           1       0.12      0.20      0.15       173

    accuracy                           0.89      3577
   macro avg       0.54      0.56      0.54      3577
weighted avg       0.92      0.89      0.90      3577

Accuracy Score:  0.8870561923399497
F1 Score:  0.14767932489451477
```

We obtained a better model performance after parameter tuning. All metrics aforementioned in the preliminary evaluation records slight uptick.

5.2. Logistic Regression

From the above 10-fold cross validation, we obtained the mean F1 score as follow:

```
Logistic Regression Mean f1 score: 0.2126537785588752
```

5.2.1. Parameter Tuning

Next, we used the package GridSearchCV again to execute an exhaustive search over parameter values.
We searched the penalty between l1 and got the best outcome at l2.
We searched the inverse of regularisation strength (C) among 0.001, 0.01, 0.1, 1, 10 and 100 got the best outcome at 0.1.

5.2.2. Result and Evaluation

Below is the preliminary result we obtained:

```
              precision    recall  f1-score   support

           0       0.98      0.75      0.85      3404
           1       0.12      0.70      0.21       173

    accuracy                           0.75      3577
   macro avg       0.55      0.72      0.53      3577
weighted avg       0.94      0.75      0.82      3577

Accuracy Score:  0.7469946882862735
F1 Score:  0.2109851787271142
```

The model gives us a normal accuracy of 0.75 and an acceptable recall. It also records a low F1 score of 0.22.

5.3. Neural Networks

To run a neural networks model, the very first thing is to import all the required packages from Tensorflow. Then we split out a validation set from the training set we had earlier for model validation use.

We started our model with 2 layers, with ReLU as the non-linear activation function. For the output layer, we use sigmoid as the non-linear activation function. It classifies prediction into 0 and 1 (binary), which is what we are looking for in "stroke". We trained our model with epochs =20 and a batch_size =512.

5.3.1. Preliminary Result and Evaluation

Result: loss: 0.3522 - accuracy: 0.9516

"Loss" is the value calculated by the loss function, representing the rate of error we seek to minimise. This model provides a decent result with an acceptable level of loss and high level of accuracy.

5.3.2. First Parameter Tuning

We still want to see if we can improve the model. This time, we added a drop-out there to help overcome overfitting. It is also good for generalization. We decided to drop out at a rate of 50% of the neurons in each hidden layer. We also added 2 extra hidden layers to see if that helps.

5.3.3. Second Result and Evaluation

Result: loss: 0.3975 - accuracy: 0.9516

Loss increases by 4.53% while accuracy remains the same.

5.3.4. Final Parameter Tuning

In the final tuning, we tried to change the number of epochs from 20 to 10.

5.3.5. Final Result and Evaluation

Result: loss: 0.6559 - accuracy: 0.7120

The result returns significant decline in both loss and accuracy. We conclude that the original neural networks model yields the best result (which is satisfying).

6. Models Comparison

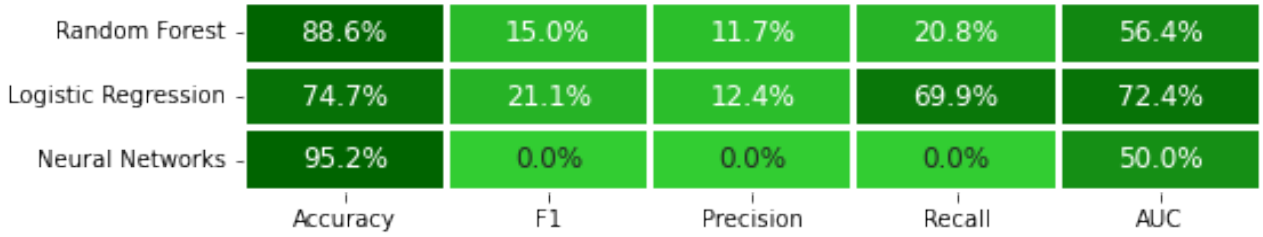

Figure 6a. Score Matrix Between Models

Neural Networks has the highest accuracy. Random Forest has a better accuracy over that of Logistic Regression, while Logistic Regression dominates in recall rate. For Neural Network, F1, precision and recall are all set to 0.0 due to no predicted samples. Based on the nature of this case, recall rate is more important because we do not want to have a high false negative rate along with a low false positive rate. So we prefer Logistic Regression Model here even it comes at the cost of a relatively small accuracy rate.

7. Conclusion

In this assignment, we looked into a stroke prediction dataset and tried to find patterns from it. As per some scholars believe, age and gender certainly play a role in determining a person at risk of stroke. Based on our findings, only some of the features are related to the risk for stroke. Other features like heart disease and average glucose level also play a role. We used three machine learning algorithms to build our models, and we tried and understood model performance can be improved after some hyper-parameter tuning. We also understood that the importance of different matrix depends on the nature of problem. For example, if we prefer correct detection of stroke, recall rate is important. Based on the result, in this case we prefer to use the Logistic Regression model the most.

8. References

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

2. Bašić Kes, V. (2016). AGE AND GENDER DIFFERENCES IN ACUTE STROKE HOSPITAL PATIENTS. Acta Clinica Croatica, 69–77. https://doi.org/10.20471/acc.2016.55.01.11

3. Shah, R. S., & Cole, J. W. (2010). Smoking and stroke: the more you smoke the more you stroke. Expert Review of Cardiovascular Therapy, 8(7), 917–932. https://doi.org/10.1586/erc.10.56