Stroke Prediction

Milestone 4: Project-3 Report

Vinay Nagaraj

DSC680, Summer 2021

Bellevue University, NE

https://vinaynagaraj88.github.io/DataScience_Portfolio/

Abstract

According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths. A stroke happens when blood stops flowing to any part of your brain, damaging brain cells. The effects of a stroke depend on the part of the brain that was damaged and the amount of damage done.

If stroke can be predicted at an early stage there is 4% lower risk of in-hospital death, 4% better odds of walking independently after leaving the hospital and also 3% better odds of being sent home instead of to an institution.

Introduction

A Report from the American Heart Association informs that an average, someone in the US has a stroke every 40 seconds. There are about 795,000 new or recurrent strokes each year, based on 1999 data. On average, someone dies of a stoke every 3 minutes and 33 seconds in the US. There are about 405 deaths from stroke each day, based on 2018 data.

In general, stroke is more likely to occur in the elderly, and stroke can lead to cerebral dysfunction such as hemiplegia, mispronunciation, and lack of consciousness. Properly managing and treating adjustable risk factors such as hypertension, smoking, diabetes, and obesity can decrease stroke occurrences. Stroke is a treatable disease, and if detected or predicted early, its severity can be greatly reduced.

Through this project, I intend to consider all the relevant information about the patient such as gender, age, various diseases, and smoking and build predictive analytics techniques that would predict the patients with high risk and is likely to get stroke. This helps in providing the advanced warning to alert the patients so that they can apply proper precautions and possibly the prevent the stroke.

Methods

The project will be carried out by utilizing the CRISP-DM model. It stands for Cross Industry Standard Process for Data Mining. The process contains 6 steps that will be followed throughout the project.

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment

As part of this paper, first 4 steps (mentioned above) will be covered in this methods section. Evaluation and deployment steps are detailed in the results and conclusion section.

- Business Understanding - Stroke is a major cause of death and functional disability globally. Recent advances in stroke care have reduced case fatality and improved functional outcomes after stroke. In particular, reperfusion therapy, such as intravenous thrombolysis and endovascular thrombectomy, significantly improves poststroke outcomes. Nevertheless, patients with stroke can still suffer from a severe disability or eventually die even after receiving appropriate treatment. An early and reliable prognosis for recovery in stroke patients is important for initiation of individual treatment and for informing patients and relatives. As part as of this project, we are trying to identify the factors that lead to stroke in patients and build a classifier model that would help hospitals in predicting if a patient is likely to suffer from stroke or not.

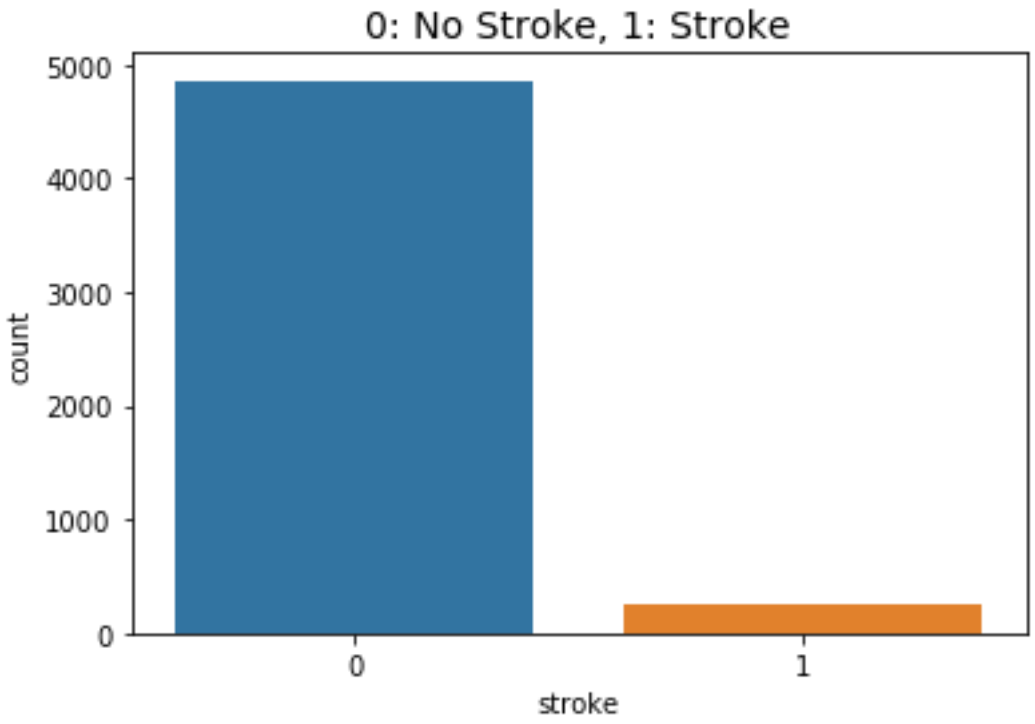
- Data Understanding - I have used the below dataset from Kaggle which contains details about patients with features such as gender, age, various diseases, and smoking status. I am planning to use this data to train the model as part of this project.

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

The dataset contains 5,110 records with 11 attributes and one target variable. Each row in the data provides relevant information about the patient. The target variable in this dataset is a binary variable reflecting the fact whether the patient had a stroke or not. “stroke”: 1 if the patient had a stroke or 0 if not.

As part of this step, we performed graph analysis to see the distribution of the variables.

Fig 1: Below plot shows the count of patients who had stroke vs No Stroke in our dataset.



Our Dataset contains a total of 4,861 rows of patients who never had stroke and 249 rows of patients who had stroke. We can observe that our dataset is highly imbalanced and we will handle that by over-sampling (SMOTE) before we perform model analysis.

Fig 2: Below plots shows the Stroke distribution by age. Age is a big factor in stroke patients - the older you get the more at risk you are.

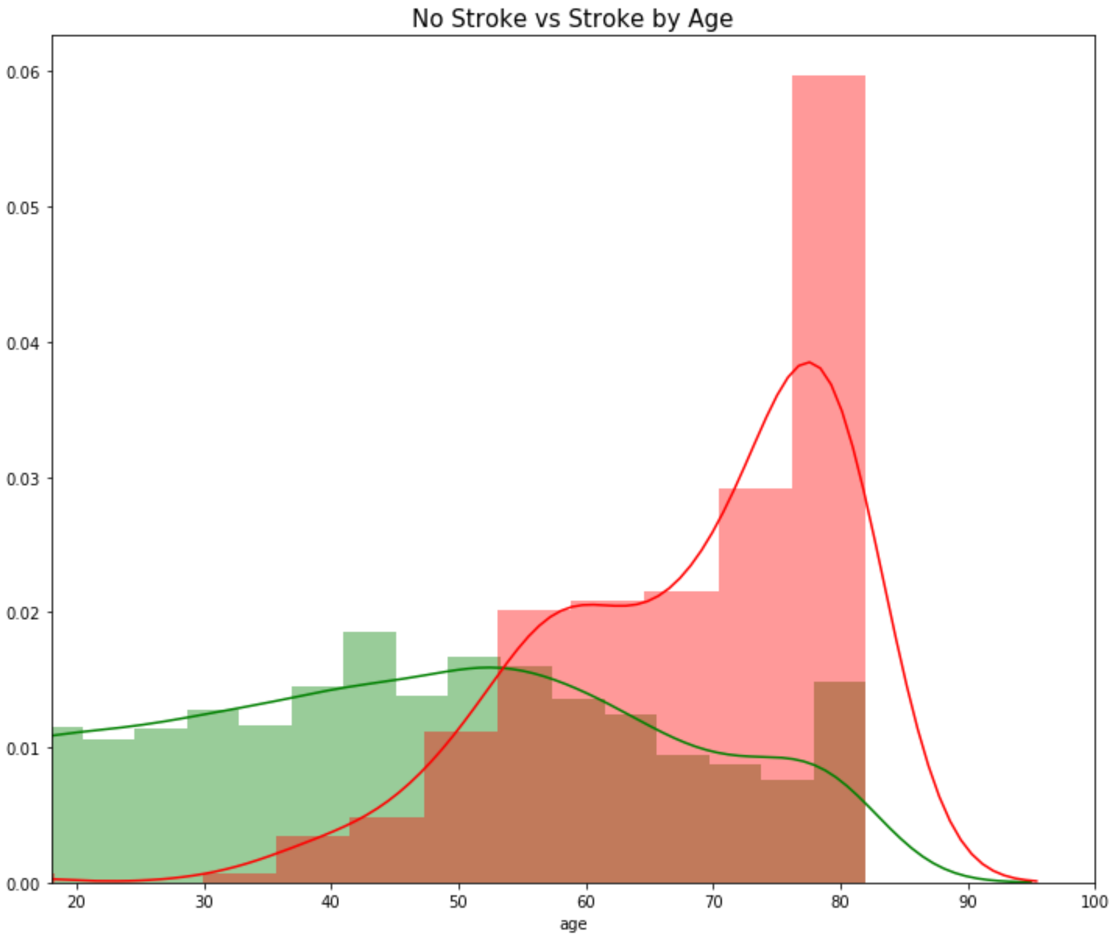


Fig 3: Below plots shows the effect of Smoking on Stroke. Being a smoker or a formerly smoker increases your risk of having a stroke. Also, looks like people who used to smoke are more prone to a Stroke than people still smoking.

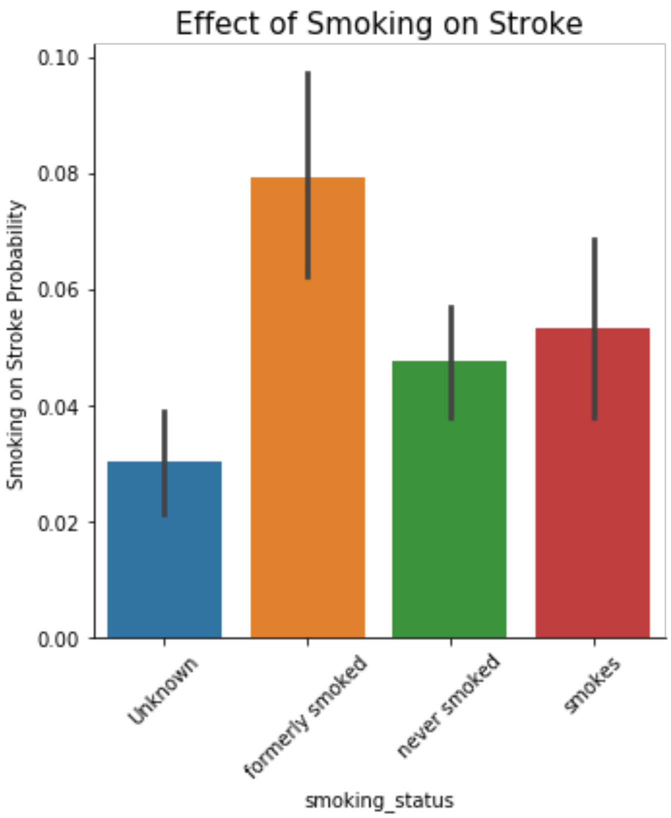


Fig 4: Below plots shows the effect of marriage on Stroke. Being married increases you chance of Stroke.

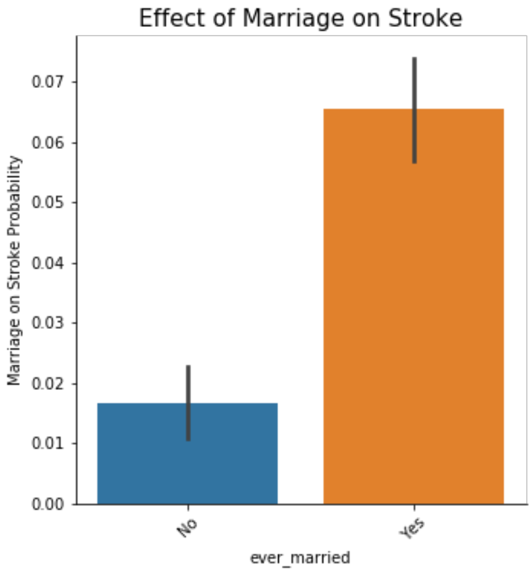


Fig 5: Below plots shows effect of Work Type on Stroke. Private work type exposes you to more stroke, than being self-employed or Govt work.

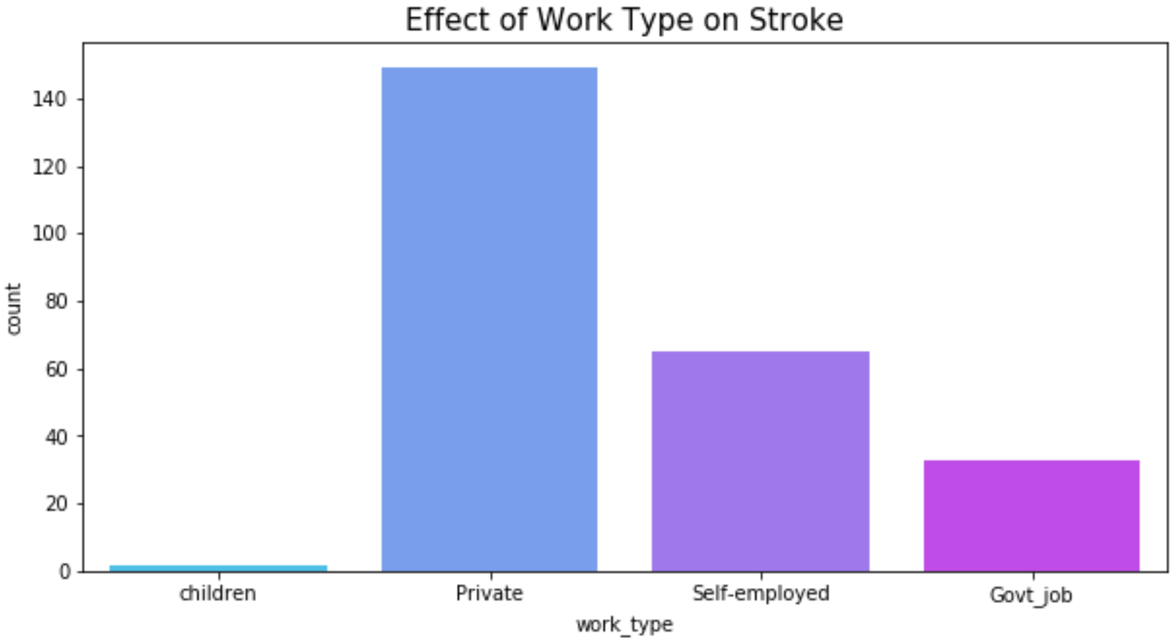


Fig 6: Below plots shows the effect of heart disease on Stroke. People with a history of heart disease are more prone to Stroke.

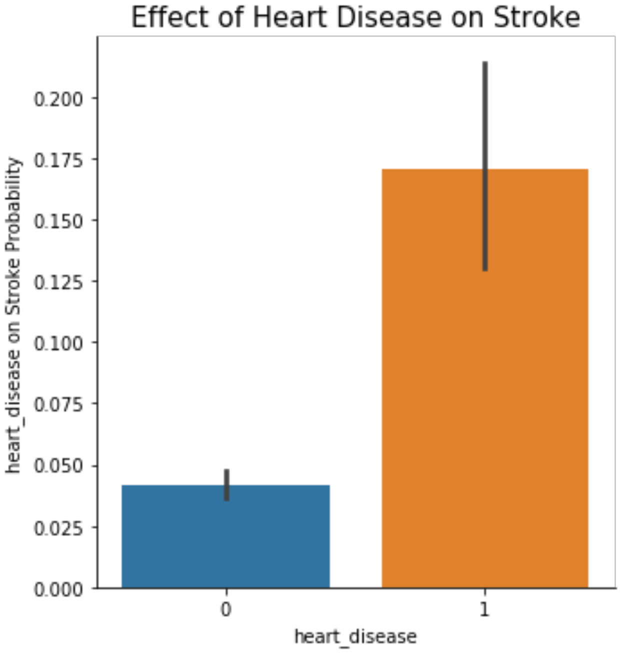


Fig 7: Below plots shows the effect of Hypertension on Stroke. People with a history of hypertension are more prone to Stroke.

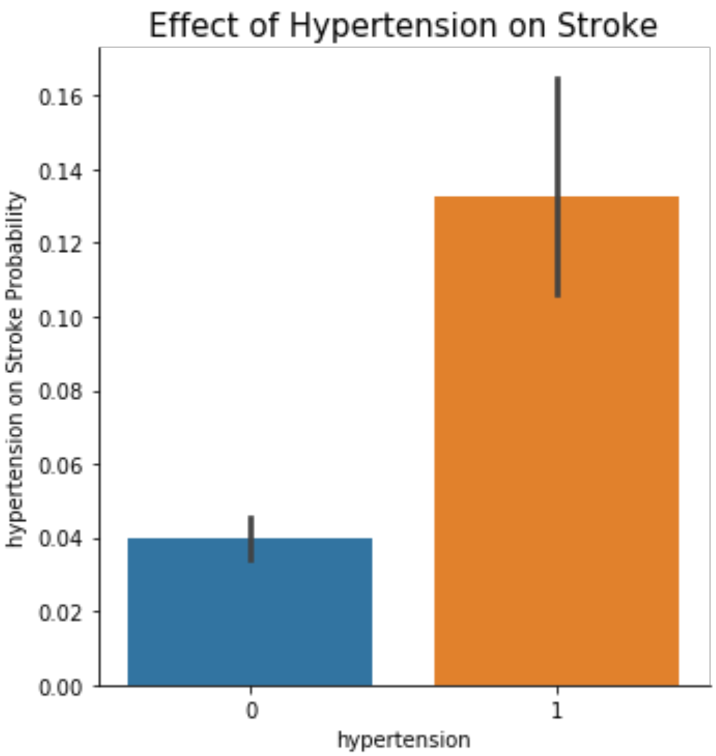


Fig 8: Below plots shows the effect of Residence Type on Stroke. People staying in Urban areas are more prone to Stroke.

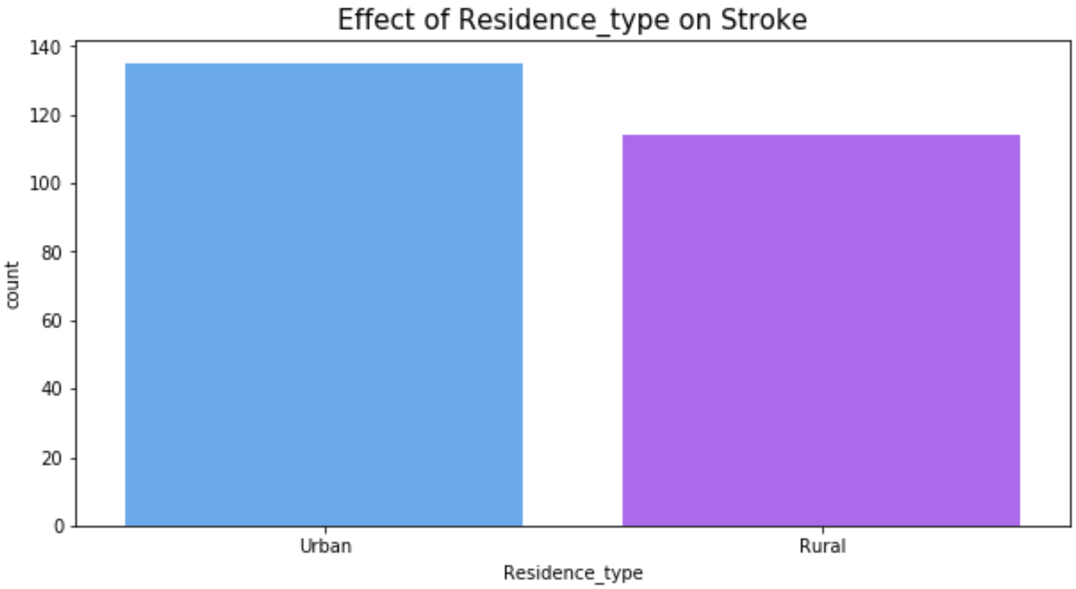
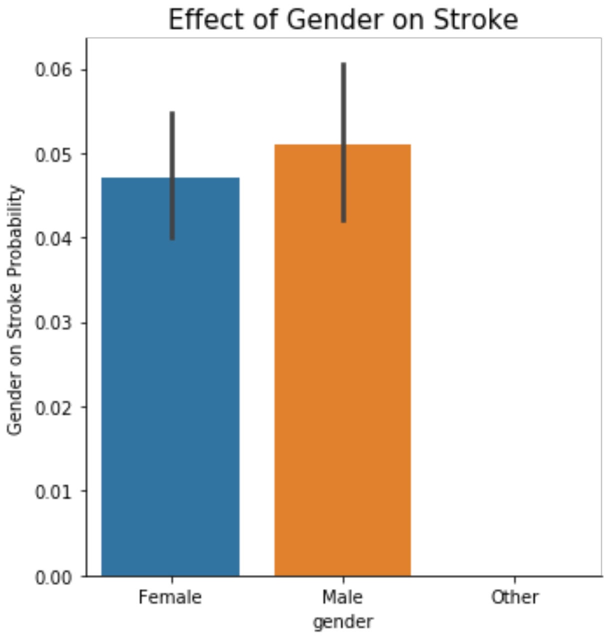


Fig 9: Below plots shows the effect of Gender on Stroke. Male are more prone to Stroke when compared to Females.



- Data Preparation – The dataset I downloaded from Kaggle was almost a clean dataset. “bmi” column had 201 missing rows which I filled using the Forward Fill technique. ‘id’ column was irrelevant to our project requirement and hence was dropped. As part of the data preparation step, I understood the characteristics of each feature in my dataset and also in preparation for modeling, I modified all the categorical features to contain numeric values as they were also important features which increased the accuracy of the classification models.
- Modeling – As part of modeling, we did notice that this dataset is severely imbalanced as most of the records belonged to no-stroke patients. So, the algorithms are much likely to classify new observations to the majority class and

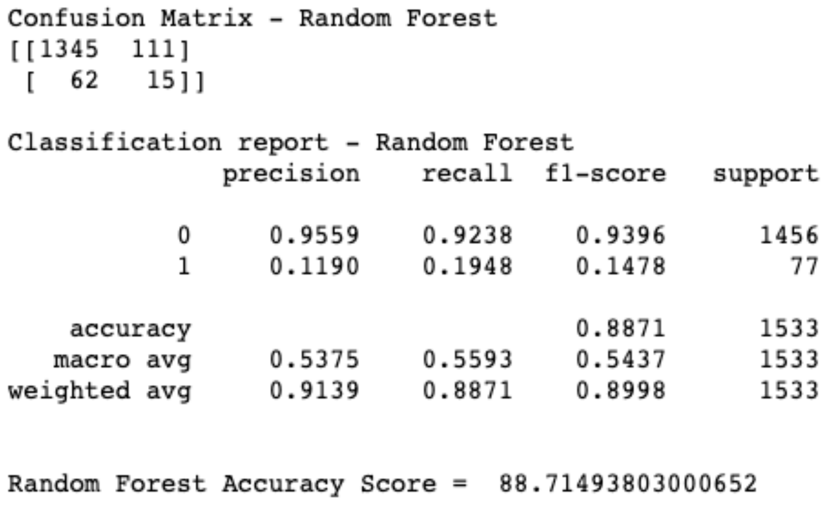
high accuracy won't tell us anything. To address the problem of imbalanced dataset, we choose to use oversampling data approach technique. Oversampling increases the number of minority class members in the training set. In order to make our data set balanced, we are using a type of oversampling called SMOTE (Synthetic Minority Oversampling Technique) and by doing that we are not losing any information from the original training set as all the observations from the minority and majority classes are retained. SMOTE works by utilizing a k-nearest neighbor algorithm to create synthetic data. Using the output data from SMOTE, I used the SelectKBest technique to understand the importance of the features in our dataset. The SelectKBest scores of all the features of our dataset can be found in Appendix A and the plot that shows the features in the order of their importance can be found in Appendix B.

Modeling

Using the data from our SMOTE results, I ran it against the Random Forest Model, k-Nearest Neighbor and Decision Tree Classifier to find out its accuracy and also plot its confusion matrix.

1. **Random forest model** creates a set of decision trees from randomly selected subset of training set. It then aggregates the votes from different decision trees to decide the final class of the test object. It provides us a high true prediction values for our dataset. We tested this model on the test dataset.

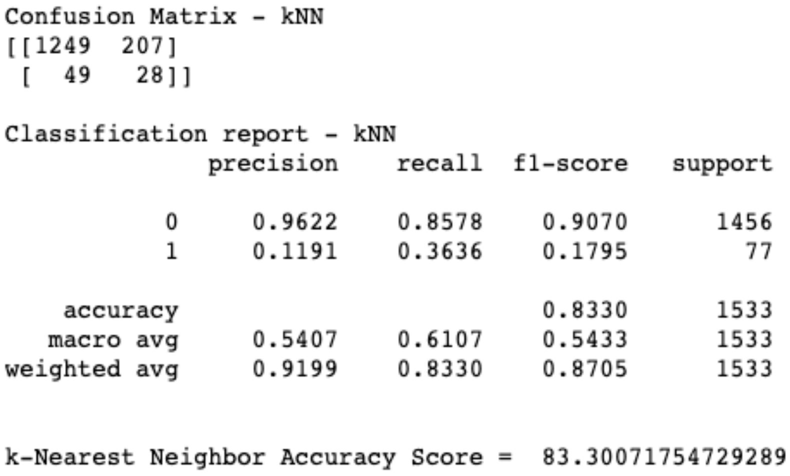
This model has an accuracy of 88.71%. 1,345 cases are true negative, meaning they are non-stroke patients and model correctly predicted them as non-stroke patients. 15 cases are true positive, meaning they are stroke patients and the model correctly predicted them as stroke patients. 62 instances are false negative and 111 are false positives.



```
Confusion Matrix - Random Forest
[[1345  111]
 [  62   15]]

Classification report - Random Forest
              precision    recall  f1-score   support

           0     0.9559    0.9238    0.9396      1456
           1     0.1190    0.1948    0.1478        77

    accuracy                         0.8871      1533
   macro avg     0.5375    0.5593    0.5437      1533
weighted avg     0.9139    0.8871    0.8998      1533


Random Forest Accuracy Score =  88.71493803000652
```

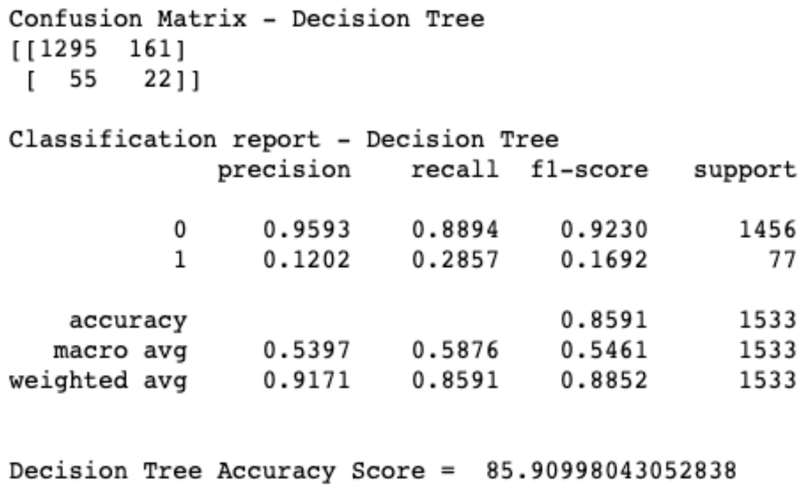
2. **K-Nearest Neighbor model** is a type of classification where the function is only approximated locally and all computation is deferred until function evaluation. Since this algorithm relies on distance for classification, if the features represent different physical units or come in vastly different scales then normalizing the training data can improve its accuracy dramatically. This model has an accuracy of 83.3%. 1,249 cases are true negative, meaning they are non-stroke patients and model correctly predicted them as non-stroke patients. 28 cases are true positive, meaning they are stroke patients and the model correctly predicted

them as stroke patients. 49 instances are false negative and 207 are false positives.



```
Confusion Matrix - kNN
[[1249  207]
 [  49   28]]

Classification report - kNN
              precision    recall  f1-score   support

           0     0.9622    0.8578    0.9070      1456
           1     0.1191    0.3636    0.1795        77

    accuracy                         0.8330      1533
   macro avg     0.5407    0.6107    0.5433      1533
weighted avg     0.9199    0.8330    0.8705      1533


k-Nearest Neighbor Accuracy Score =  83.30071754729289
```

3. **Decision Tree model** is a type of Supervised Machine Learning where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves. This model has an accuracy of 85.9%. 1,295 cases are true negative, meaning they are non-stroke patients and model correctly predicted them as non-stroke patients. 22 cases are true positive, meaning they are stroke patients and the model correctly predicted them as stroke patients. 55 instances are false negative and 161 are false positives.



```
Confusion Matrix - Decision Tree
[[1295  161]
 [  55   22]]

Classification report - Decision Tree
              precision    recall  f1-score   support

           0     0.9593    0.8894    0.9230      1456
           1     0.1202    0.2857    0.1692        77

    accuracy                         0.8591      1533
   macro avg     0.5397    0.5876    0.5461      1533
weighted avg     0.9171    0.8591    0.8852      1533


Decision Tree Accuracy Score =  85.90998043052838
```

Results

Detecting Stroke at an early stage and providing medication for that is very crucial to a patient's health condition. Warning signs of an ischemic stroke may be evident as early as seven days before an attack and require urgent treatment to prevent serious damage to the brain. Early treatment results in a greater chance of recovery, a reduced likelihood of permanent disability and lesser need for extensive rehabilitation.

- We see that age, ever_married, smoking_status are the most important features when it comes to predicting stroke-prone individuals, based on the current dataset.
- Age is an important factor in Stroke patients. It is necessary to take care of our health as we grow older and make sure not to miss out on our annual checkups.
- Smoking is injurious to health and can increase the chances of Stroke
- People staying in Urban areas are more prone to Stroke when compared to Rural areas.

- SMOTE Technique used to overcome imbalanced data.
- Among all the algorithms we used Random Forest was best suited with accuracy of 88.71%

Acknowledgements

I’m indebted to the communities behind the multiple open-source software packages on which we depend. I would like to thank my family for their understanding of our time in this endeavor. Last by not least, thanks to Prof Catie Williams and my classmates for their guidance and feedback.

References

[1] Fedesoriano. (2021, January 26). Stroke prediction dataset. Kaggle. https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

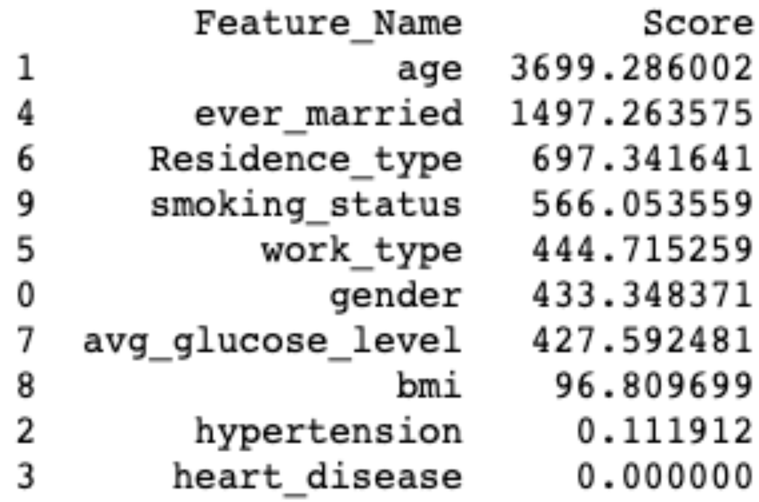
[2] Lin, S. (2021, March 10). Stroke prediction. Medium. https://medium.com/geekculture/stroke-prediction-d26c15f9d1

[3] Ding, L., Lingling Ding Department of Neurology, Liu, C., Chelsea Liu https://orcid.org/0000-0001-9942-9955 Department of Epidemiology, Li, Z., Zixiao Li Correspondence to: Zixiao Li, Wang, Y., Wang, Y. W. Y., For Sources of Funding and Disclosures, &amp; Al., E. (2020, October 27). Incorporating artificial intelligence into stroke care and research. Stroke. https://www.ahajournals.org/doi/10.1161/STROKEAHA.120.031295

[4] How artificial intelligence can predict and detect stroke. DAIC. (2019, May 17). https://www.dicardiology.com/article/how-artificial-intelligence-can-predict-and-detect-stroke

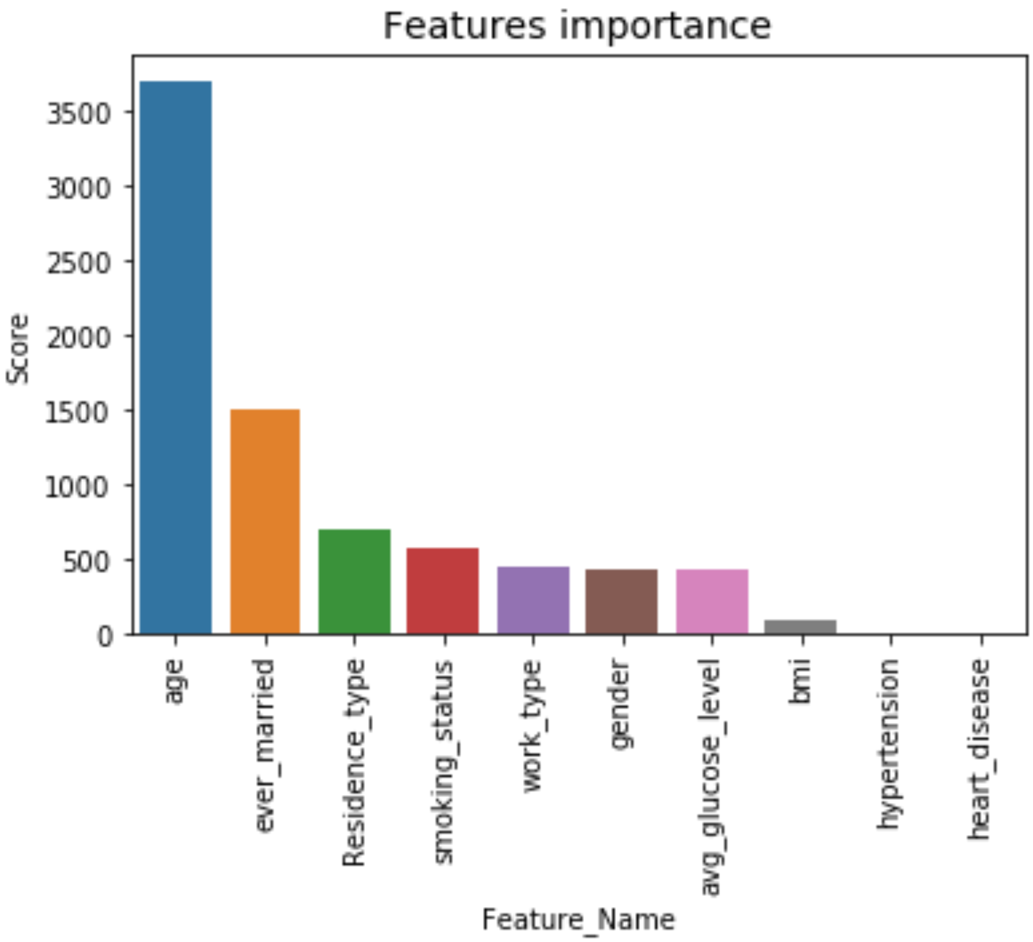
Appendix

Appendix A:



```
        Feature_Name        Score
1                age  3699.286002
4       ever_married  1497.263575
6     Residence_type   697.341641
9     smoking_status   566.053559
5          work_type   444.715259
0             gender   433.348371
7  avg_glucose_level   427.592481
8                bmi    96.809699
2       hypertension     0.111912
3      heart_disease     0.000000
```

Appendix B:


Features importance
Score
3500
3000
2500
2000
1500
1000
500
0
age
ever_married
Residence_type
smoking_status
work_type
gender
avg_glucose_level
bmi
hypertension
heart_disease
Feature_Name