Heart Stroke Prediction Model

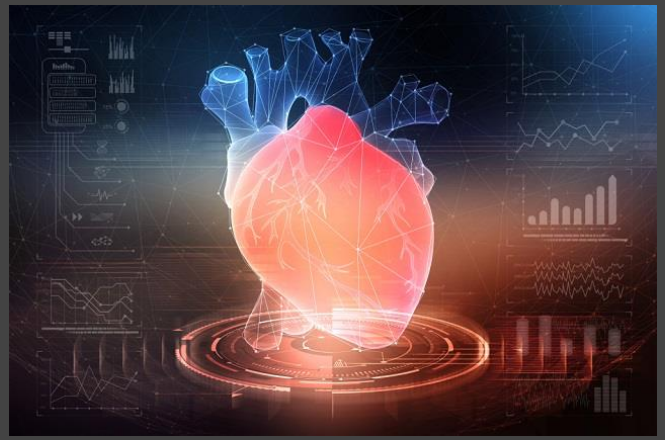
Created By
Ritesh Rajurkar

Topic

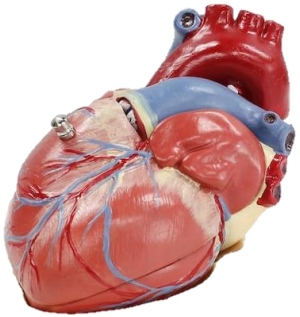
- Problem Statement
- Data Summary
- Analysis (EDA)
- Challenges
- Conclusion

Problem Statement

The objective of this study is to construct a prediction model for predicting stroke and to assess the accuracy of the model. We will explore multiple different models to see which produces reliable and repeatable results



What is Heart Stroke?



“A Heart Stroke happens when blood stops flowing to any part of your brain, damaging brain cells. The effect of a stroke depend on the part of the brain that was damaged and the amount of damage done. Knowing how your brain works can help you understand your stroke.”

Data Summary

This dataset is used to predict whether a patient is likely to get stroke based on the input parameters like gender, age, various diseases, and smoking status. Each row in the data provides relevant information about the patient.

This Dataset contain 5110 rows and 11 columns/Features such as Id, gender, age, hypertension, heart_disease, ever_married, work_type, residence_type, ave_glucose_level, bmi, smoking status, stroke.

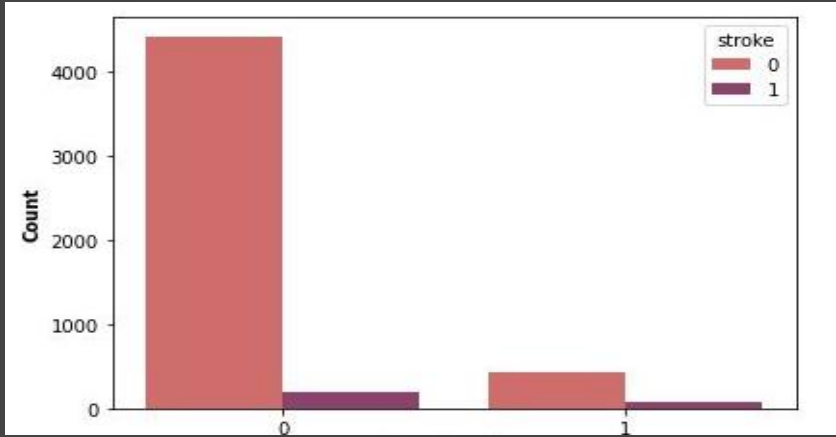
	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1

Overview of Data Attributes

1) **Id:** unique identifier
2) **Gender:** "Male", "Female" or "Other"
3) **Age:** age of the patient
4) **Hypertension:** 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
5) **Heart_disease:** 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
6) **Ever_married:** "No" or "Yes"
7) **Work_type:** "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
8) **Residence_type:** "Rural" or "Urban"
9) **Avg_glucose_level:** average glucose level in blood
10) **Bmi:** body mass index
11) **Smoking_status:** "formerly smoked", "never smoked", "smokes" or "Unknown"*
12) **Stroke:** 1 if the patient had a stroke or 0 if not

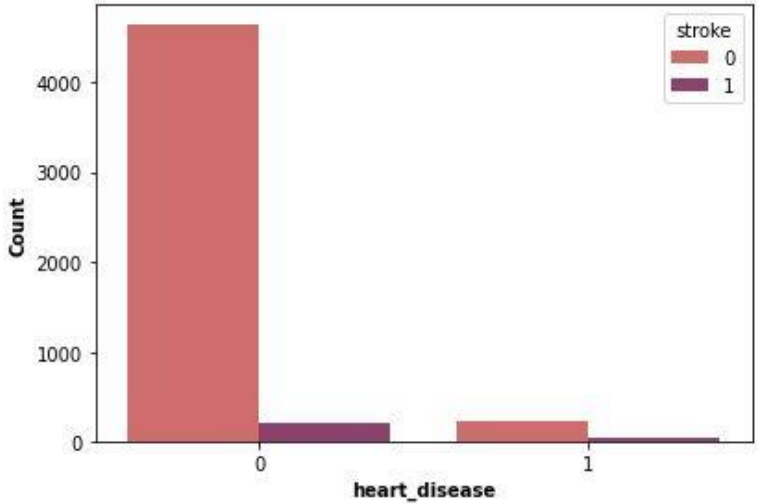
Note: "Unknown" in smoking_status means that the information is unavailable for this patient

Hypertension Analysis



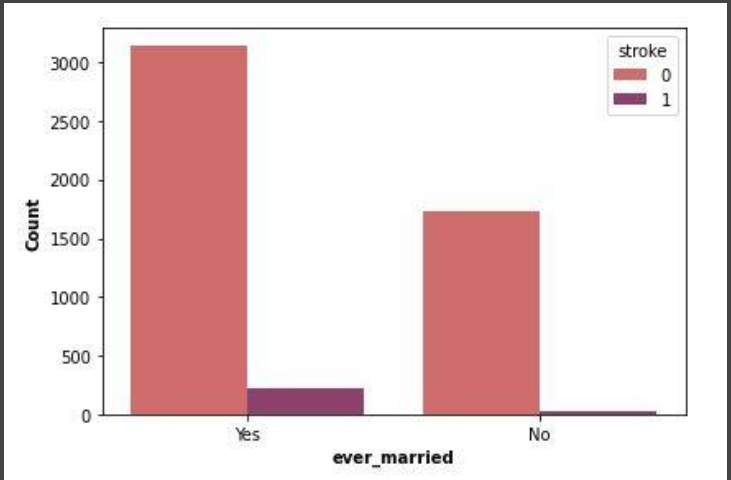
Hypertension: Subjects that previously diagnosed with hypertension have highly risk of having stroke.

Heart Disease Analysis



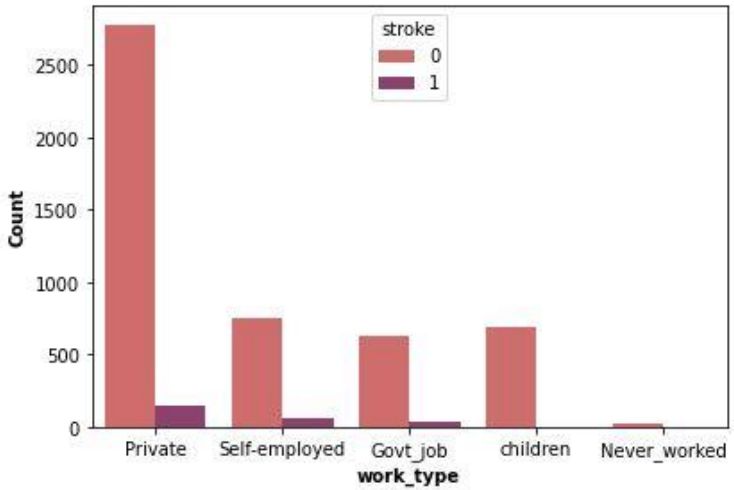
Heart disease: Subjects that previously diagnosed with heart disease have highly risk of having stroke.

Ever married Analysis



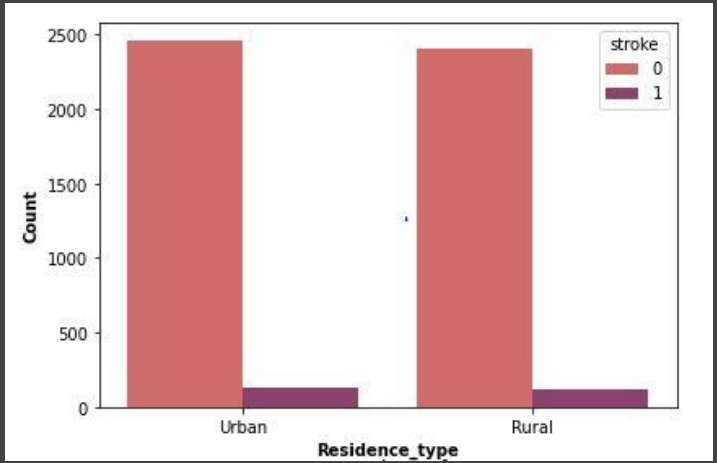
Ever married: **Subjects that ever married have highly risk of having stroke.**

Work Type Analysis



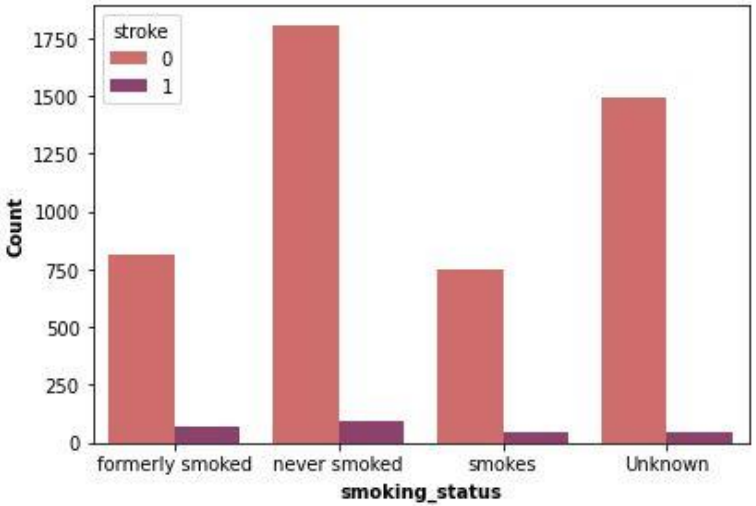
Work type: **Subjects that have any work experience and in government related work have highly risk of having stroke while those with no work experience barely experienced stroke**

Residence type Analysis



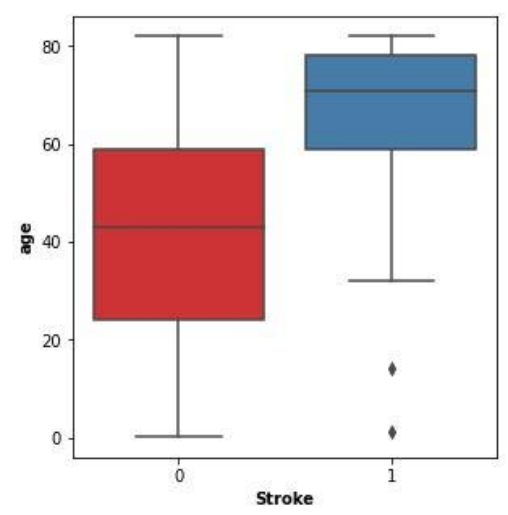
Residence type: **No obvious relationship with likelihood of experiencing stroke.**

Smoking Status Analysis



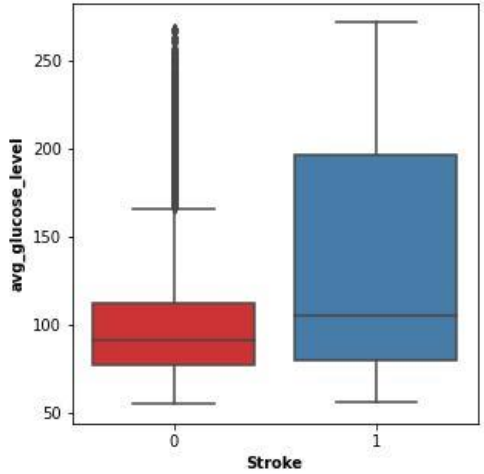
Smoking status: Being a smoker or former smoker increases risk of having a stroke.

Age wise Analysis



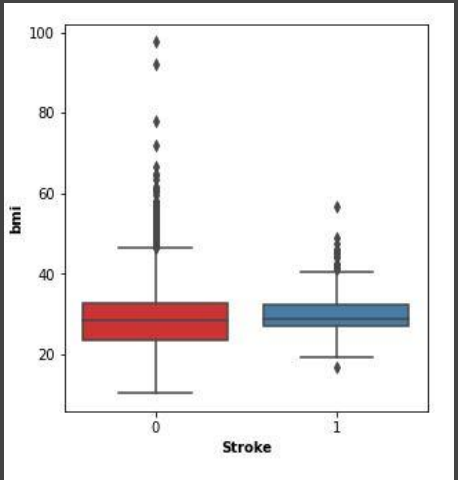
Age: Subjects with stroke tends to have higher mean age.

Ave glucose level Analysis



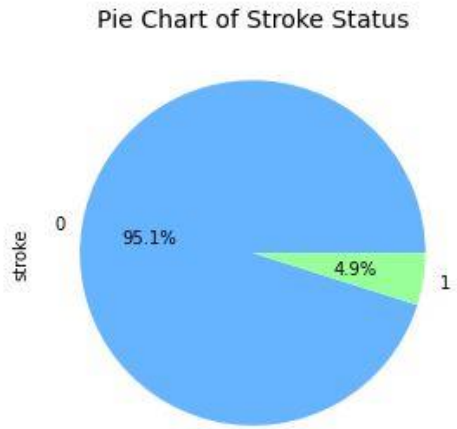
Ave glucose level: Subjects with stroke tends to have higher average glucose level.

BMI Analysis



BMI: bmi index does not give much indication on the likelihood of experiencing stroke. bmi index for super obesity is 50. Outliers in this feature should be replaced to its highest limit (50)

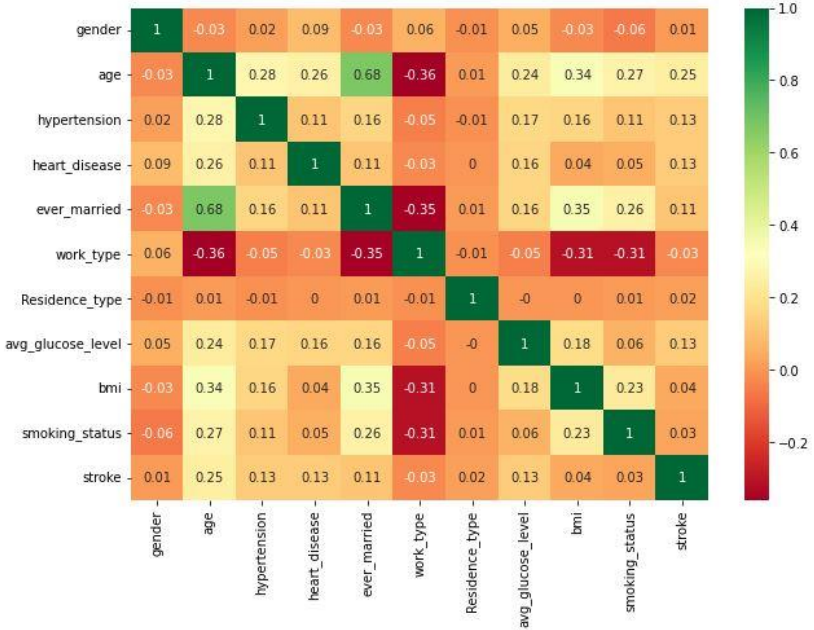
Stroke Status Analysis



Stroke Status: **4.9% of the population in this dataset is diagnosed with stroke**

Correlation Heatmap

From the correlation matrix, we can verify the presence of multicollinearity between some of the variables. For instance, the ever_married and age column has a correlation of 0.68. Between this two attributes, age contains more information on whether one is susceptible to stroke.



Test Set Prediction

```
--------------------Logreg--------------------
[[1464    0]
 [  69    0]]
              precision    recall  f1-score   support

           0       0.95      1.00      0.98      1464
           1       0.00      0.00      0.00        69

    accuracy                           0.95      1533
   macro avg       0.48      0.50      0.49      1533
weighted avg       0.91      0.95      0.93      1533
```

```
--------------------Random Forest--------------------
[[1462    2]
 [  69    0]]
              precision    recall  f1-score   support

           0       0.95      1.00      0.98      1464
           1       0.00      0.00      0.00        69

    accuracy                           0.95      1533
   macro avg       0.48      0.50      0.49      1533
weighted avg       0.91      0.95      0.93      1533
```

```
--------------------Support Vector Machine------------
[[1412   52]
 [  64    5]]
              precision    recall  f1-score   support

           0       0.96      0.96      0.96      1464
           1       0.09      0.07      0.08        69

    accuracy                           0.92      1533
   macro avg       0.52      0.52      0.52      1533
weighted avg       0.92      0.92      0.92      1533
```

```
--------------------Decision Tree--------------------
[[1394   70]
 [  64    5]]
              precision    recall  f1-score   support

           0       0.96      0.95      0.95      1464
           1       0.07      0.07      0.07        69

    accuracy                           0.91      1533
   macro avg       0.51      0.51      0.51      1533
weighted avg       0.92      0.91      0.91      1533
```

```
--------------------kNN--------------------
[[1457    7]
 [  66    3]]
              precision    recall  f1-score   support

           0       0.96      1.00      0.98      1464
           1       0.30      0.04      0.08        69
```

Take note that recall can be thought of as a measure of a classifiers completeness. A low recall for stroke (1) indicates many False Negatives.

Sum of Accuracy Score

From the accuracy summary, **Logistic Regression, Random Forest and KNN models** all gives high accuracy score of **0.95**. However, it is also important to consider the error type and recall value of each model. Models with **0.95** accuracy score generally have high false negative as shown in the confusion matrix. High false negative indicates **type 2 error**. For our study on stroke prediction, we want to avoid **type 2 error** as it means that we fail to identify subjects that has **stroke** and deem them stroke free instead. Inspecting from the classification report above, **Naive Bayes Model** has fit our objective although the accuracy is **0.87**.

```
Summary of Accuracy Score


Decision Tree Model:  0.9126
Logreg Model:  0.955
Random Forest Model:  0.9537
Support Vector Machine Model:  0.9243
kNN Model:  0.9524
Naive Bayes Model:  0.8728
KMeans Model:  0.7834
```

Conclusion

- Various model was used to predict whether a person is subjected to stroke. Naive Bayes model yields a very good performance as indicated by the model accuracy which was found to be 87.28%.
- Using the mean cross-validation, we can conclude that we expect the model to be around 87.31% accurate on average.
- If we look at all the 10 scores produced by the 10-fold cross-validation, we can also conclude that there is a relatively small variance in the accuracy between folds, hence the model is independent of the particular folds used for training.
- Our original model accuracy is 87.28% and the mean cross-validation accuracy is 87.31%. Thus, the 10-fold cross-validation accuracy does result in performance improvement for this model.
- Naive Bayes model can be further improve by tuning hyperparameters to get the better result or adjusting the probablity threshold to improve its performance.