The Chances of a Person Having a Stroke

What is a Stroke?

A Stroke is defined as 'a sudden disabling attack or loss of consciousness caused by an interruption in the flow of blood to the brain, especially through thrombosis.'

Stroke is a leading cause of death and disability, causing around 38,000 deaths each year in the UK.

There are around 100,000 strokes every year in the UK
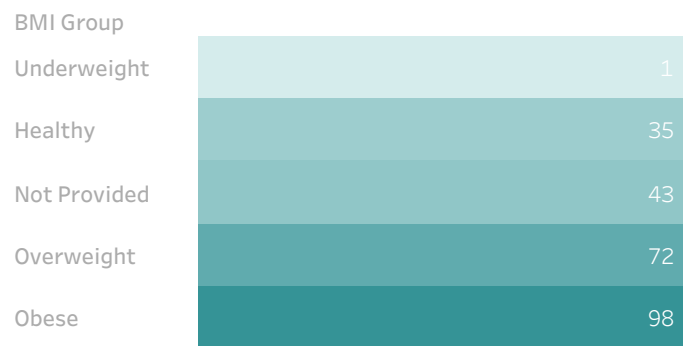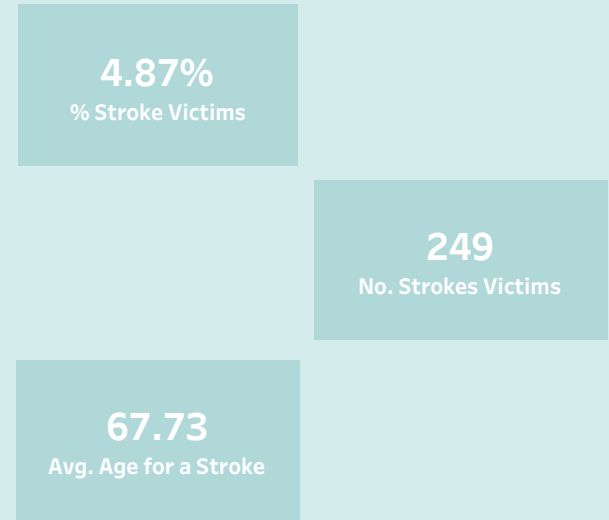
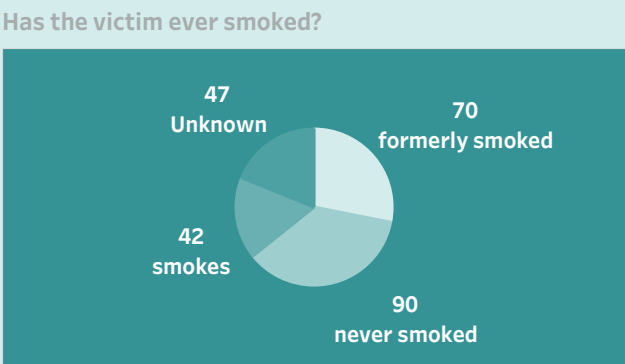126,000 hospital admissions in England per year

In the UK there are approximately 1.3 million people living with stroke

Presented by -
Hanna, Issa, Vera and Violetta

The Chances of a Person Having a Stroke
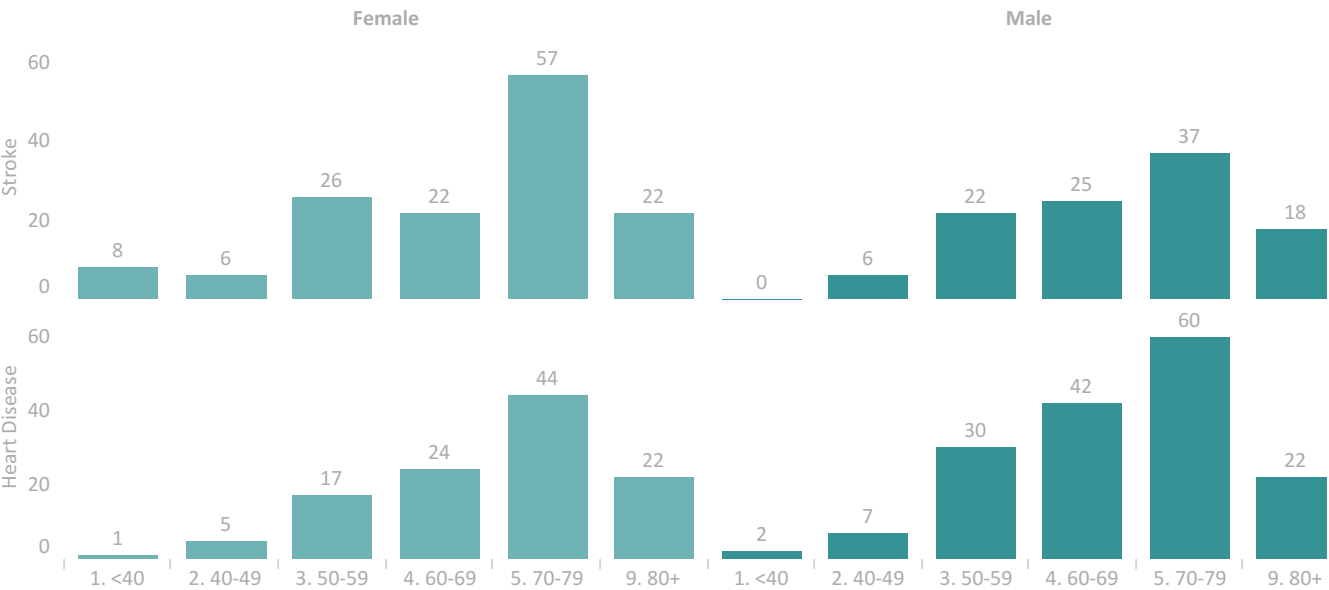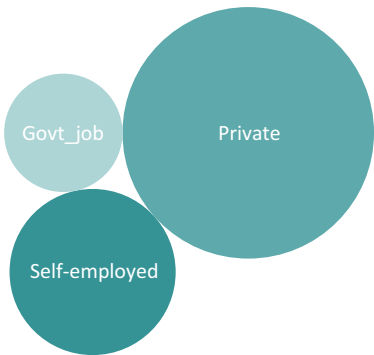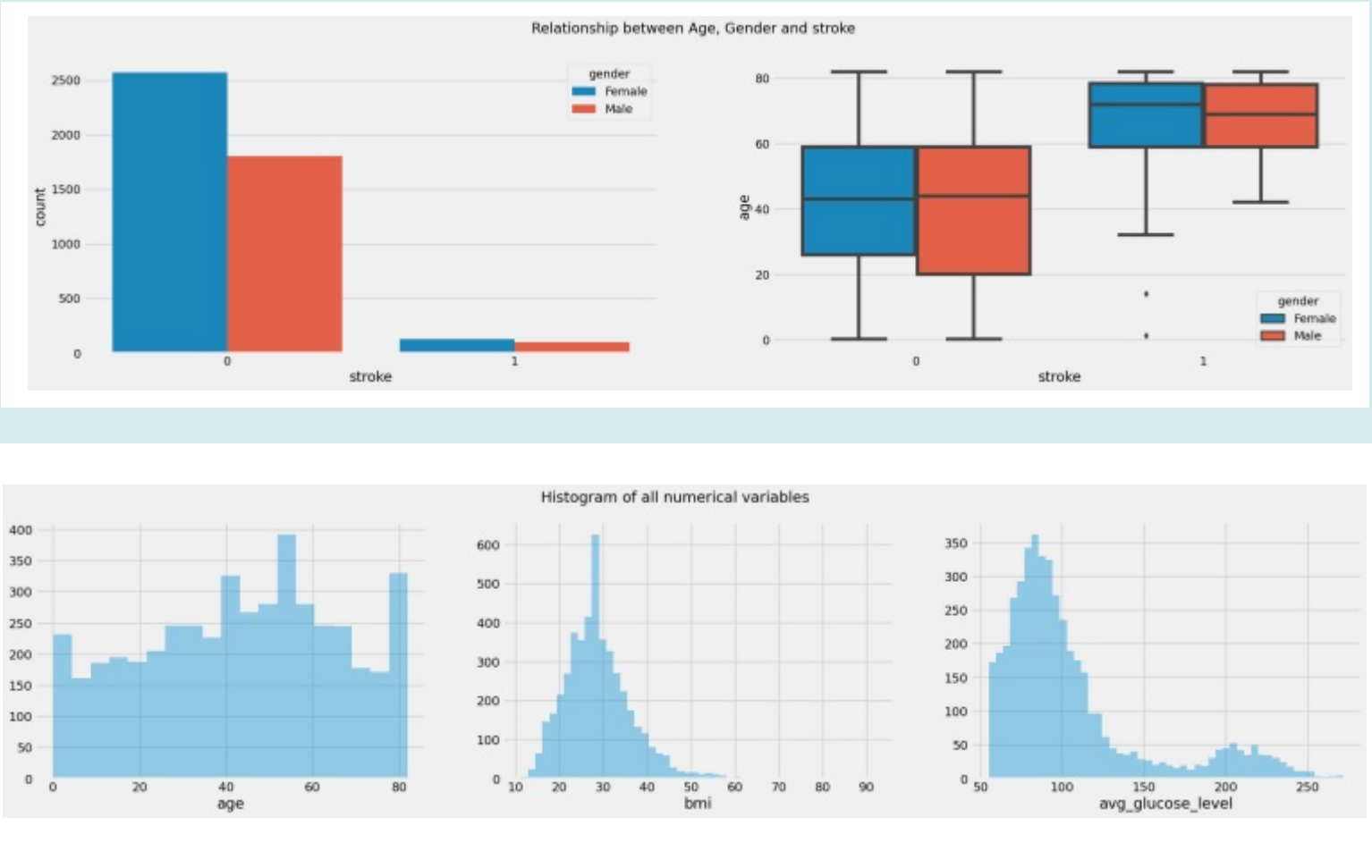
Stroke Statistics

Has the victim ever smoked?

- 47 Unknown
- 70 formerly smoked
- 42 smokes
- 90 never smoked

4.87% % Stroke Victims

249 No. Strokes Victims

67.73 Avg. Age for a Stroke

BMI Group

BMI Group	
Underweight	1
Healthy	35
Not Provided	43
Overweight	72
Obese	98

Work Type of Stroke Victims

- Govt_job
- Private
- Self-employed

Female — Stroke

1. <40	2. 40-49	3. 50-59	4. 60-69	5. 70-79	9. 80+
8	6	26	22	57	22

Male — Stroke

1. <40	2. 40-49	3. 50-59	4. 60-69	5. 70-79	9. 80+
0	6	22	25	37	18

Female — Heart Disease

1. <40	2. 40-49	3. 50-59	4. 60-69	5. 70-79	9. 80+
1	5	17	24	44	22

Male — Heart Disease

1. <40	2. 40-49	3. 50-59	4. 60-69	5. 70-79	9. 80+
2	7	30	42	60	22

The Chances of a Person Having a Stroke

Understanding the Data - Part 1

Relationship between Age, Gender and stroke

Histogram of all numerical variables
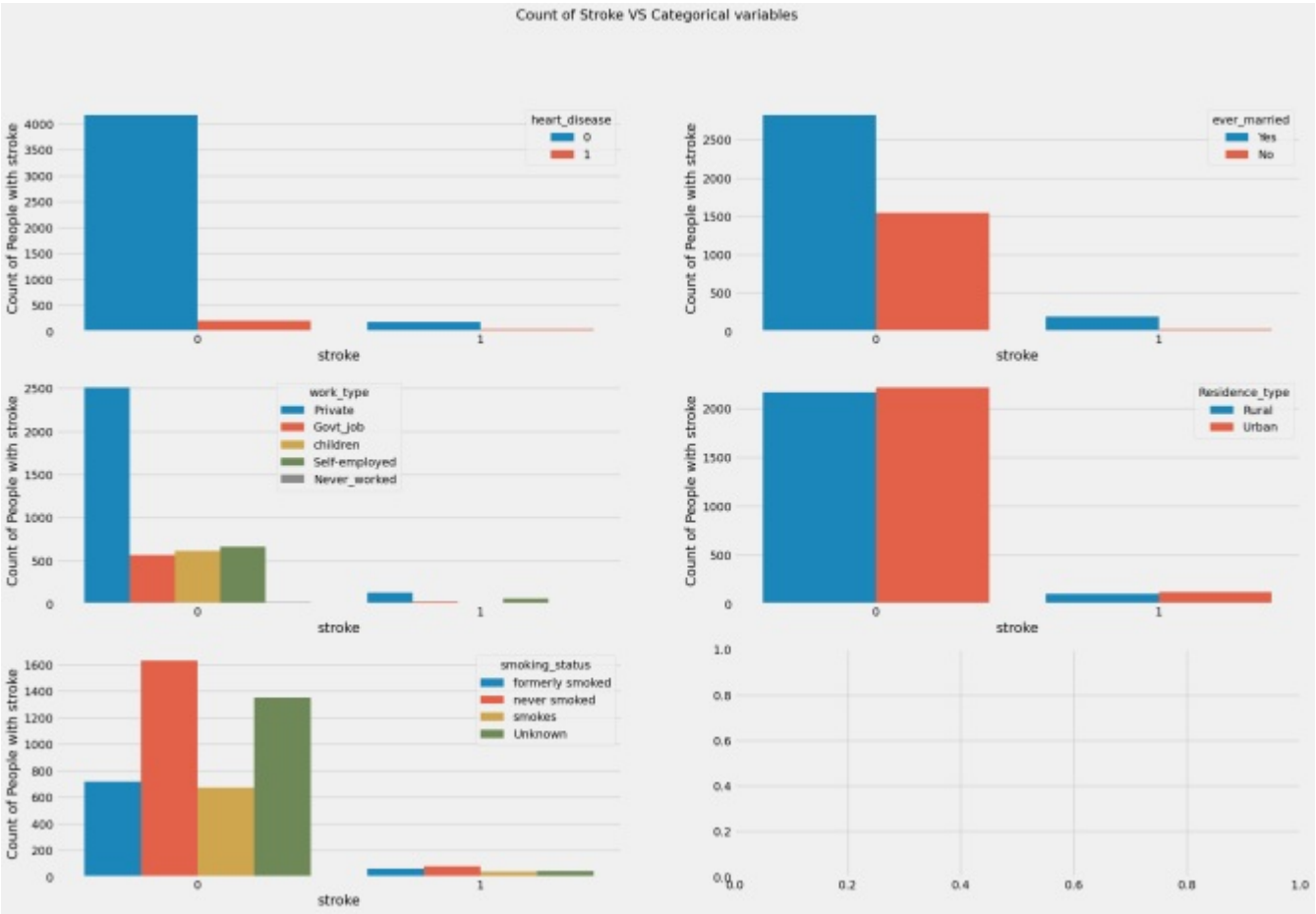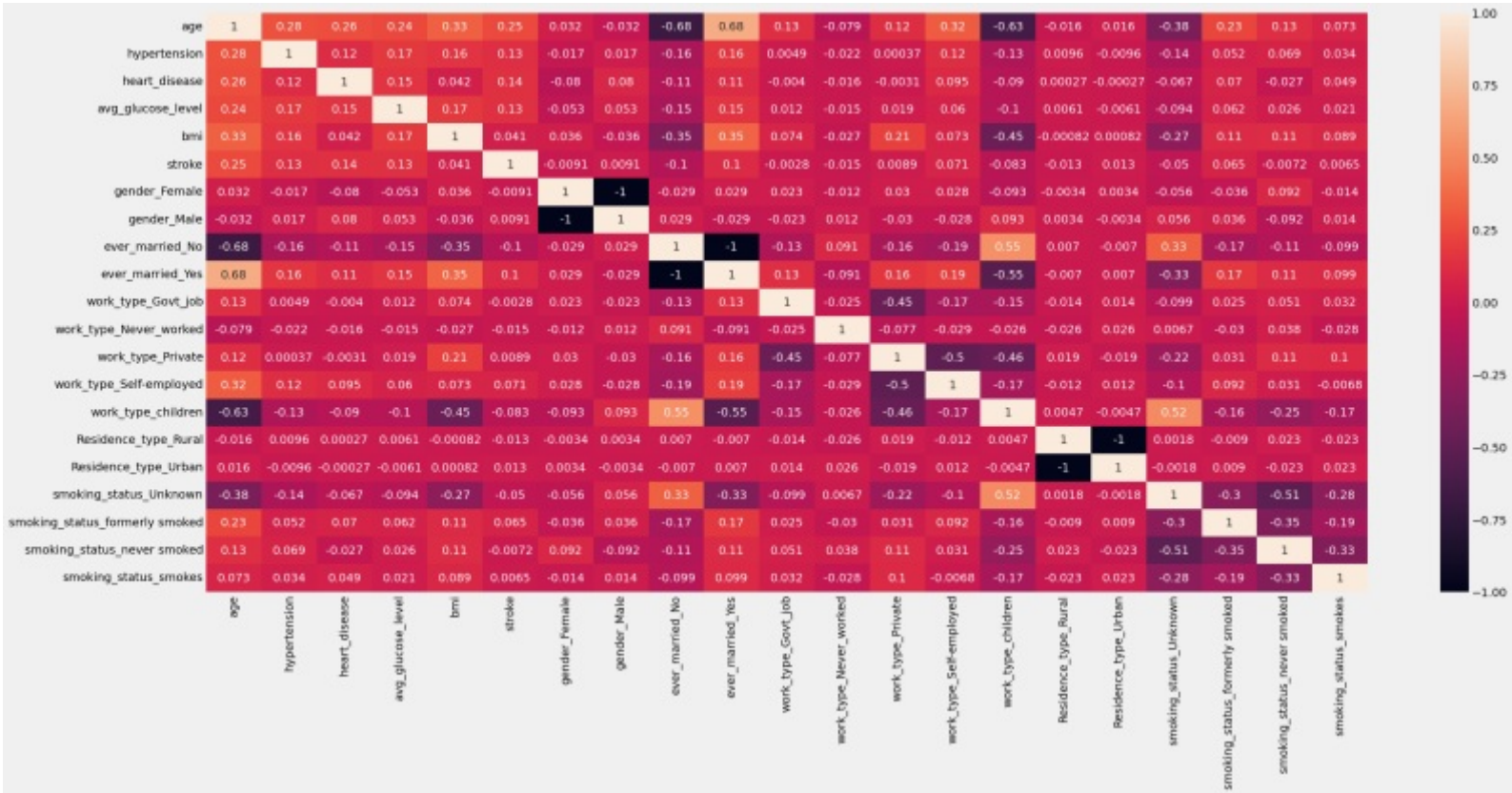
The Chances of a Person Having a Stroke

Understanding the Data - Part 2

The Chances of a Person Having a Stroke
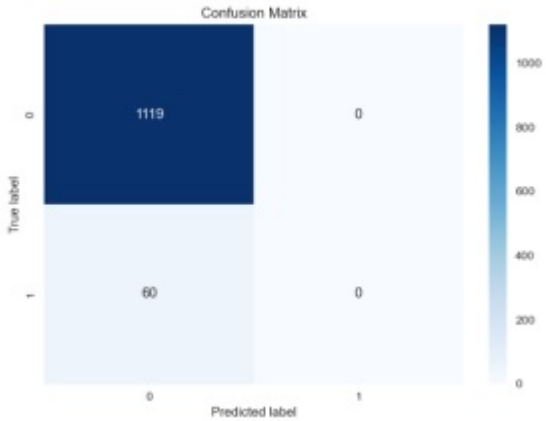
Understanding the Data - Part 3
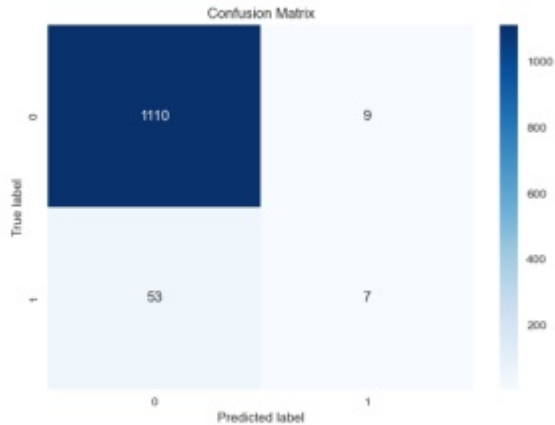
The Chances of a Person Having a Stroke

Machine Learning

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
dummy	Dummy Classifier	0.9558	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	0.1390
lightgbm	Light Gradient Boosting Machine	0.9445	0.7813	0.1643	0.3022	0.2030	0.1775	0.1905	0.2400
rf	Random Forest Classifier	0.9424	0.7860	0.1643	0.2587	0.1991	0.1713	0.1767	0.2440
et	Extra Trees Classifier	0.9421	0.7867	0.1300	0.2329	0.1644	0.1373	0.1445	0.2240
gbc	Gradient Boosting Classifier	0.9255	0.7955	0.2962	0.2507	0.2630	0.2253	0.2301	0.2500
dt	Decision Tree Classifier	0.9045	0.5647	0.1919	0.1234	0.1492	0.1014	0.1046	0.1350
ada	Ada Boost Classifier	0.8897	0.8010	0.4062	0.1764	0.2447	0.1953	0.2159	0.1790
knn	K Neighbors Classifier	0.8439	0.6809	0.3824	0.1189	0.1811	0.1216	0.1466	0.1370
lr	Logistic Regression	0.7697	0.8399	0.7267	0.1287	0.2185	0.1550	0.2368	0.4700
nb	Naive Bayes	0.7636	0.8204	0.7462	0.1287	0.2194	0.1556	0.2407	0.1320
qda	Quadratic Discriminant Analysis	0.7630	0.8092	0.7038	0.1226	0.2086	0.1442	0.2221	0.1390
svm	SVM - Linear Kernel	0.7594	0.0000	0.7405	0.1292	0.2189	0.1552	0.2383	0.1340
lda	Linear Discriminant Analysis	0.7536	0.8407	0.7667	0.1261	0.2165	0.1521	0.2412	0.1360
ridge	Ridge Classifier	0.7509	0.0000	0.7600	0.1239	0.2130	0.1482	0.2362	0.1310

Dummy Classifier
Confusion Matrix

	Predicted 0	Predicted 1
True label 0	1119	0
True label 1	60	0

Linear Discriminant Analysis
Confusion Matrix

	Predicted 0	Predicted 1
True label 0	1110	9
True label 1	53	7

The Chances of a Person Having a Stroke

XGBClassifier and Resampled_XGBClassifier

Confusion Matrix

Confusion Matrix
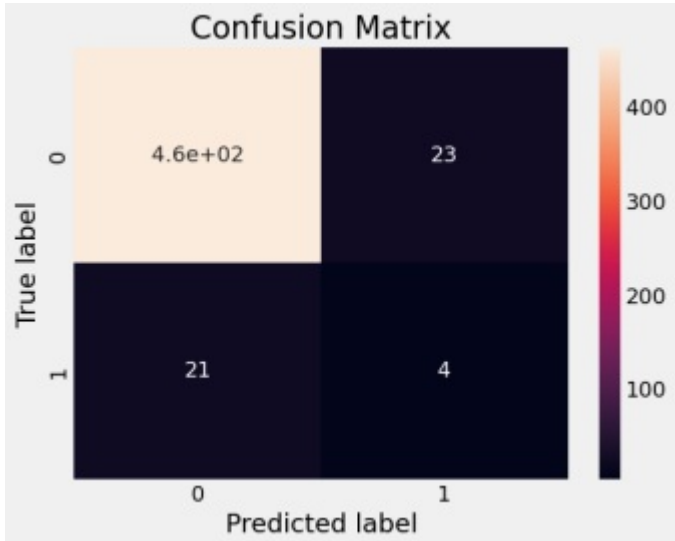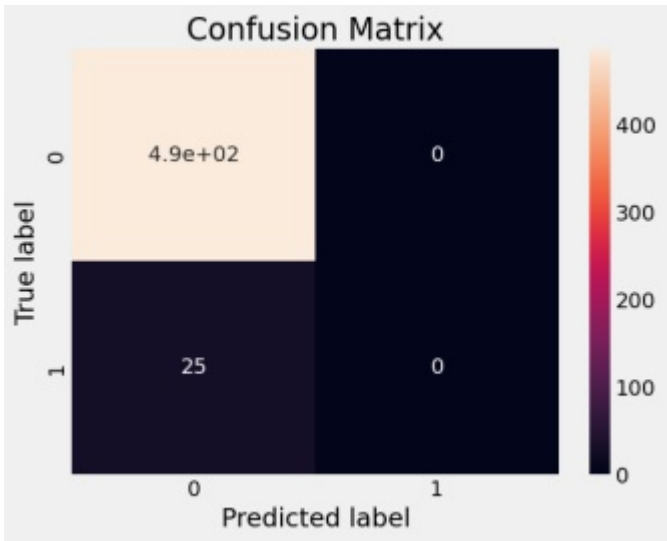
The Chances of a Person Having a Stroke

Feature Importance

Feature Importance
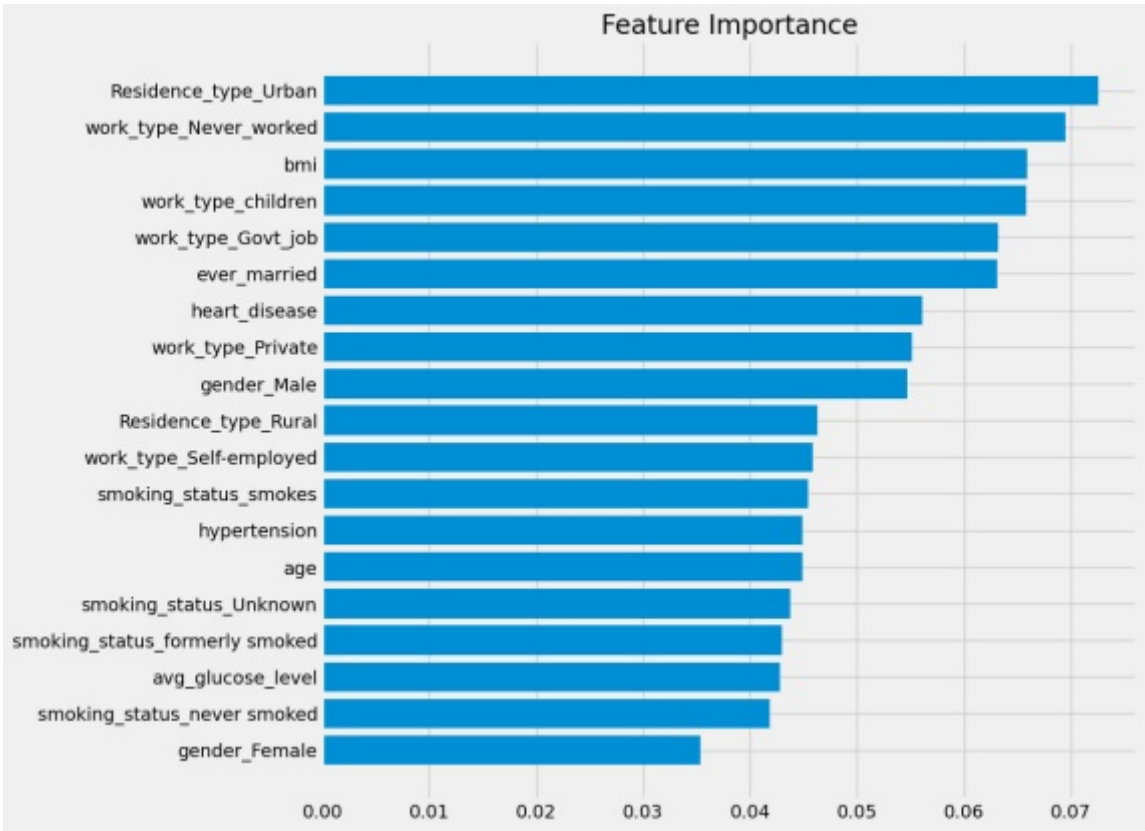
The Chances of a Person Having a Stroke

Play area

```
What is your gender?
(0 - Female, 1 - Male)
1
What is your age?
78
Do you have hypertension?
(0 - No, 1 - Yes)
1
Do you have heart_disease?
(0 - No, 1 - Yes)
1
Have you been ever married?
(0 - No, 1 - Yes)
0
What is your work type?
(0 - Private, 1 - Self employed, 2 - Children, 3 - Goverment job, 4 - Never worked)
1
What is your residence type?
(0 - Rural, 1 - Urban)
0
What is your avgerage glucose level?
92.62
What is your bmi?
40
What is your smoking habit?
(0 - Never smoked, 1 - Unknown, 2 - Formerly smoked, 3 - Smokes)
2


Stroke positive
```

The Chances of a Person Having a Stroke

Conclusion

To determine the best model, we need to consider both accuracy and the confusion matrix. Here are a few observations:

> The first two models (Dummy Classifier and LDA) have similar accuracies, with the Dummy Classifier having a slightly higher accuracy. However, both models have a significant number of false negatives (FN), indicating a relatively high misclassification rate for positive instances (stroke cases).

> The XGBClassifier (with mean BMI) achieved the highest accuracy (95.1%) among all the models. However, it also had a higher number of false negatives (FN) compared to the first two models.

> The fourth model, which used the resampled dataset with XGBClassifier (with mean BMI), had a lower accuracy (92.4%) compared to the others. It also had a considerable number of false negatives (FN) and false positives (FP)