Stroke Prediction

Areeb Shafqat

28th November 2021

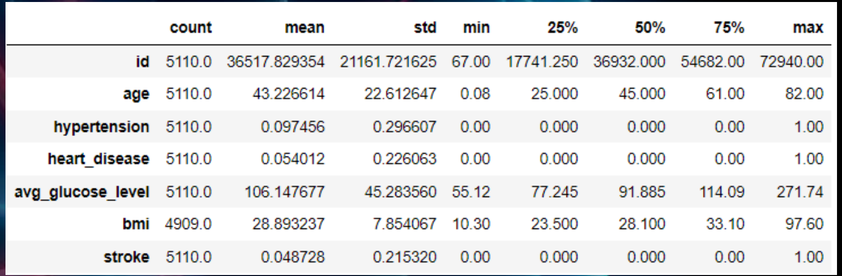
Context & Data Description

Based on reports from the World Health Organization (WHO), stroke is the 2nd leading cause of deaths worldwide and 1 in every 6 deaths resulted from a cardiovascular disease was from stroke. In the US for example, 23-29% of people have a stroke at least once and around 795,000 people per year suffer from this.

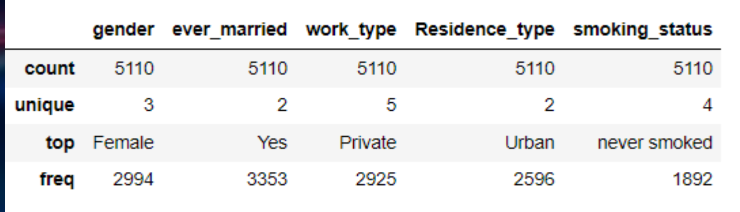
The dataset used for predicting stroke contains 5110 observations of patients either suffering from a stroke or not with the following attributes: https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

- Patient ID
- Gender (Male/Female/Other)
- Age
- Hypertension (0 – no, 1 – yes)
- Heart Disease (0 – no, 1 – yes)
- Ever Married ("Yes", "No")
- Work Type / Occupation e.g., Private
- Residence Type (Rural/Urban)
- Average Glucose Level in Blood
- Body Mass Index
- Smoking Status e.g., Smokes, Never Smoked
- Stroke (0 – no, 1 – yes)

Summary Statistics:

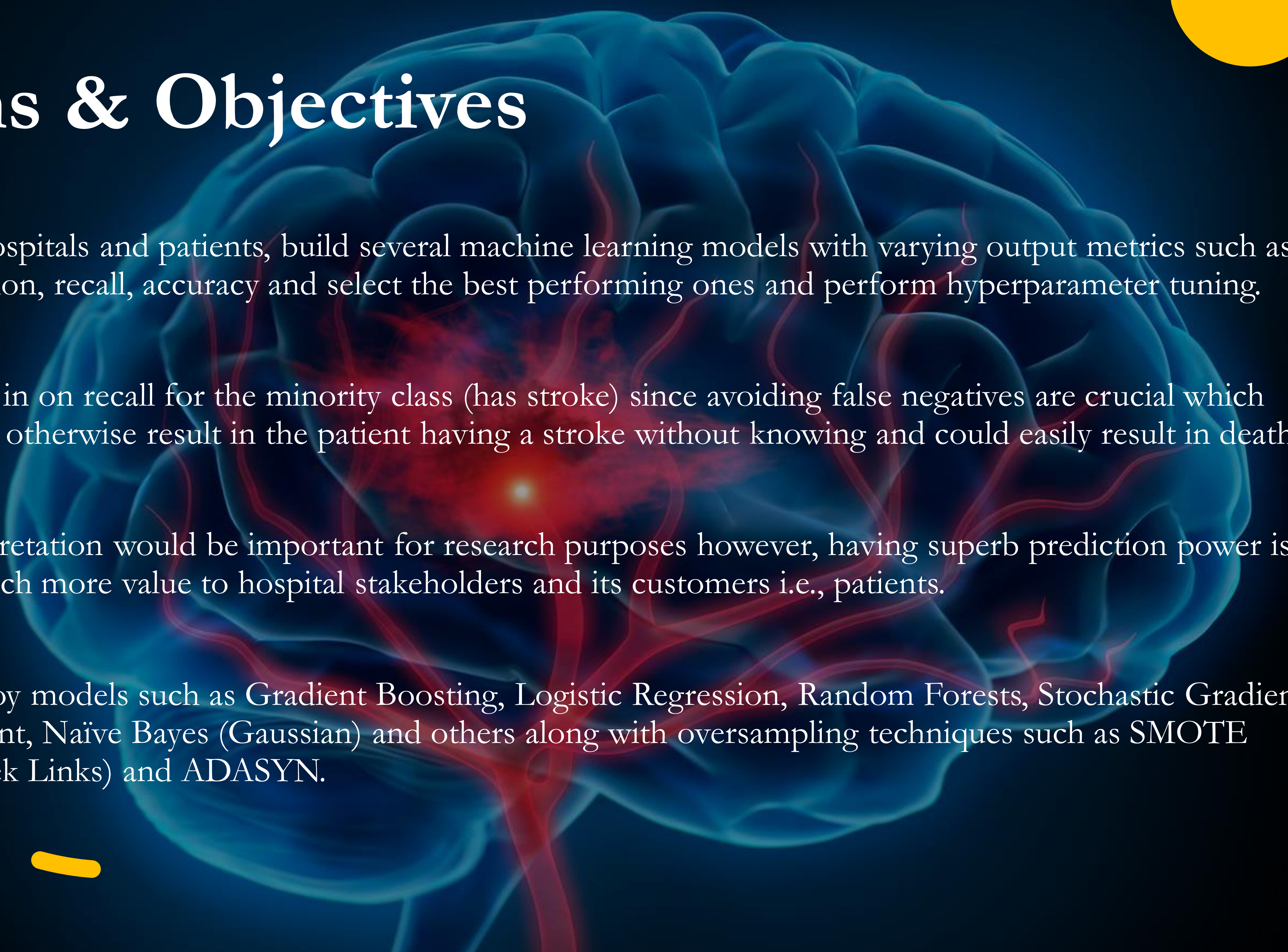


	count	mean	std	min	25%	50%	75%	max
id	5110.0	36517.829354	21161.721625	67.00	17741.250	36932.000	54682.00	72940.00
age	5110.0	43.226614	22.612647	0.08	25.000	45.000	61.00	82.00
hypertension	5110.0	0.097456	0.296607	0.00	0.000	0.000	0.00	1.00
heart_disease	5110.0	0.054012	0.226063	0.00	0.000	0.000	0.00	1.00
avg_glucose_level	5110.0	106.147677	45.283560	55.12	77.245	91.885	114.09	271.74
bmi	4909.0	28.893237	7.854067	10.30	23.500	28.100	33.10	97.60
stroke	5110.0	0.048728	0.215320	0.00	0.000	0.000	0.00	1.00



	gender	ever_married	work_type	Residence_type	smoking_status
count	5110	5110	5110	5110	5110
unique	3	2	5	2	4
top	Female	Yes	Private	Urban	never smoked
freq	2994	3353	2925	2596	1892

Aims & Objectives



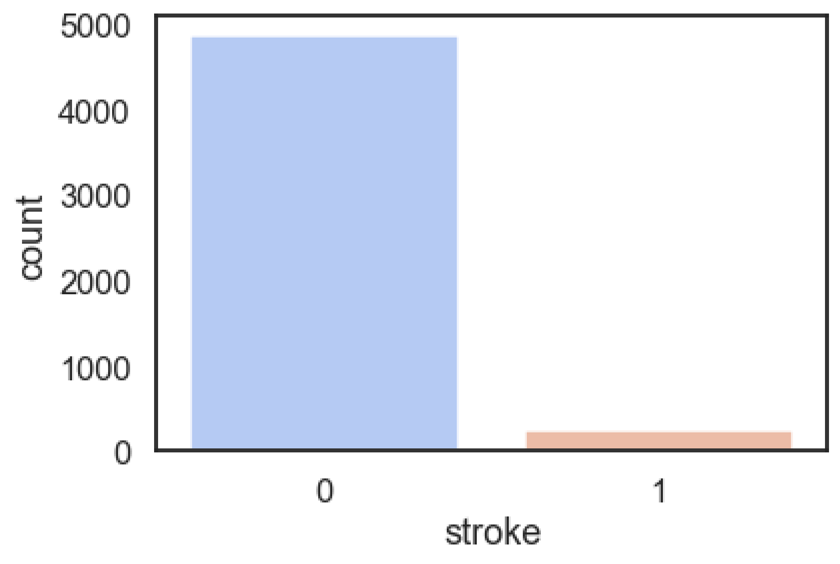
- ✓ For hospitals and patients, build several machine learning models with varying output metrics such as precision, recall, accuracy and select the best performing ones and perform hyperparameter tuning.

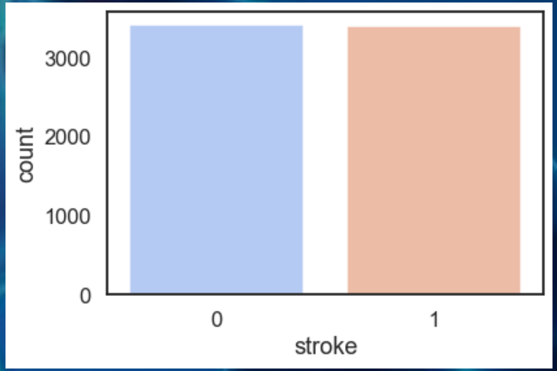
- ✓ Focus in on recall for the minority class (has stroke) since avoiding false negatives are crucial which would otherwise result in the patient having a stroke without knowing and could easily result in death.

- ✓ Interpretation would be important for research purposes however, having superb prediction power is of much more value to hospital stakeholders and its customers i.e., patients.

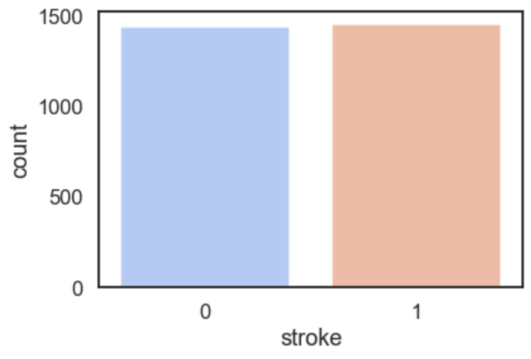
- ✓ Employ models such as Gradient Boosting, Logistic Regression, Random Forests, Stochastic Gradient Descent, Naïve Bayes (Gaussian) and others along with oversampling techniques such as SMOTE (Tomek Links) and ADASYN.

Data Engineering and Analysis





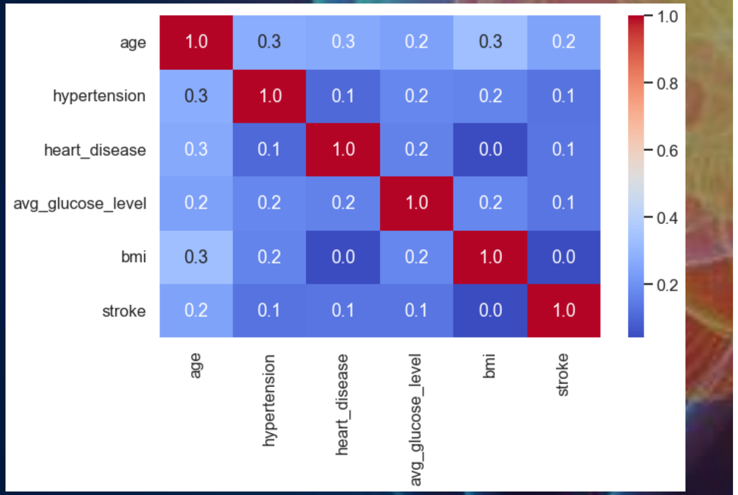
Y Train ADASYN

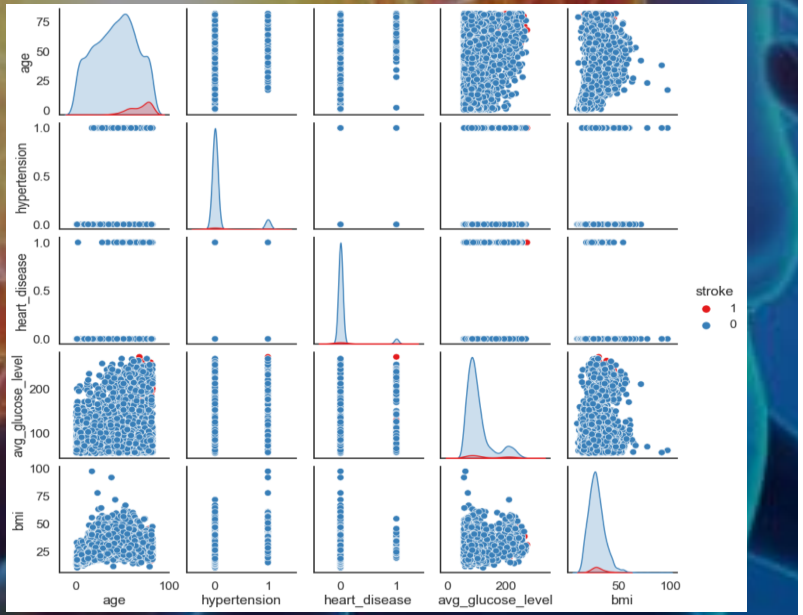


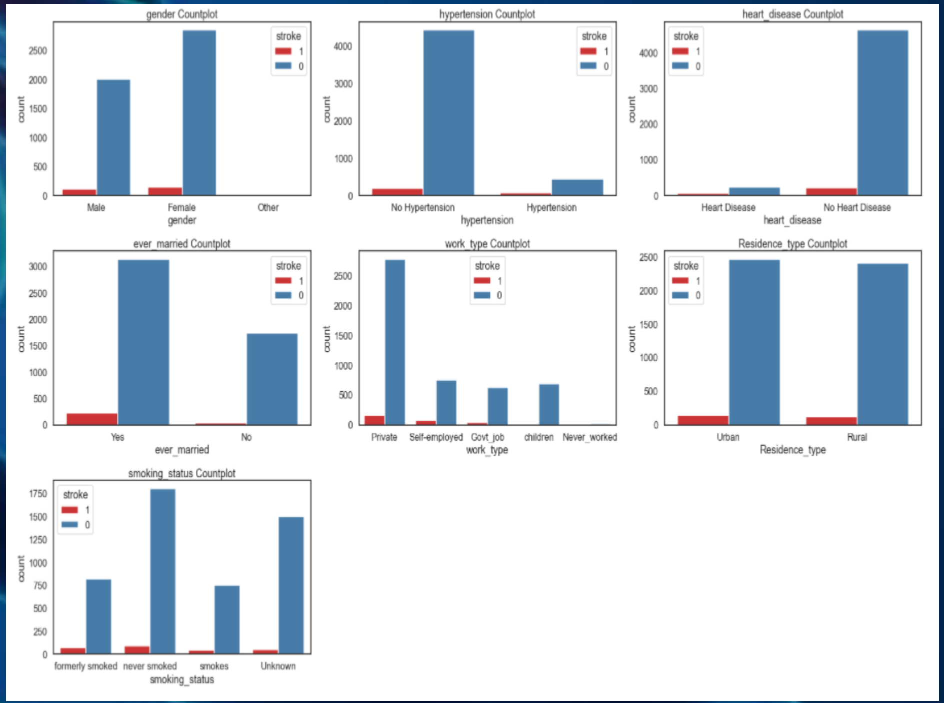
Y Test ADASYN

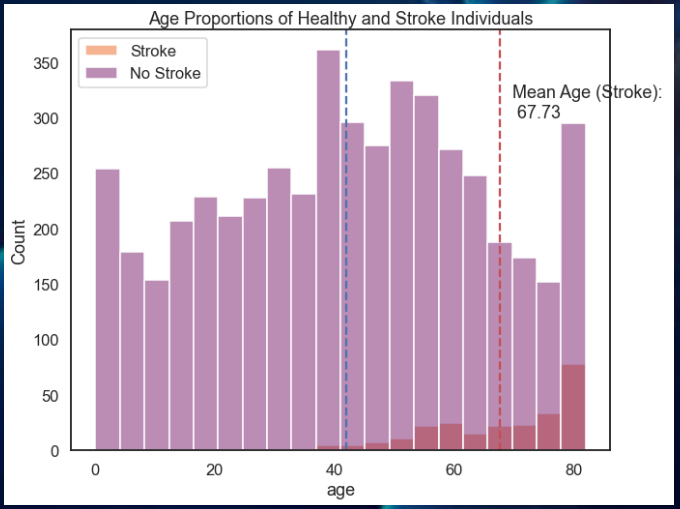
Huge class imbalance requiring oversampling/under sampling techniques. We will use oversampling ADASYN and SMOTE Tomek Links after performing same train test splits where SMOTE Tomek is a mix of over and under sampling. ADASYN provided the best results regarding recall and f1-scores and therefore will only be mentioned in this report.

Data Engineering and Analysis





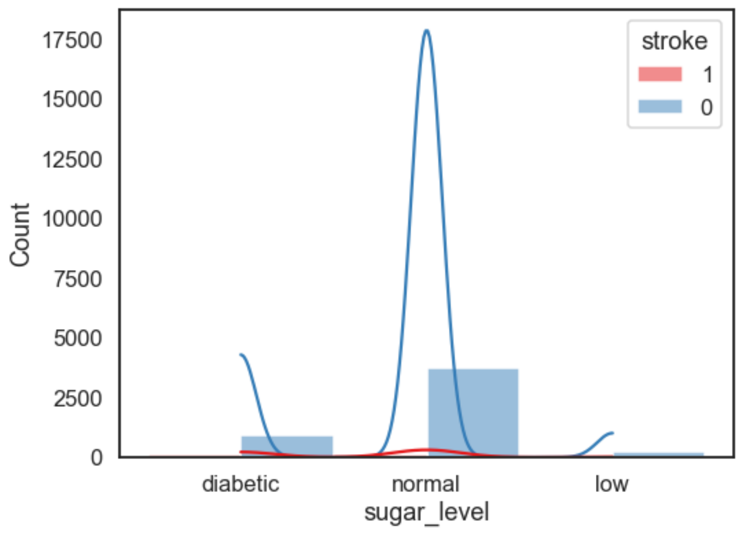


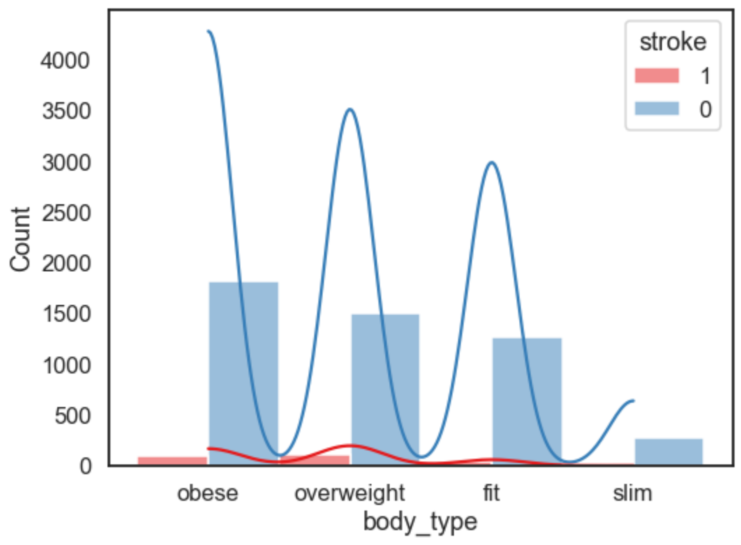


- No significant correlations exist amongst variables for multicollinearity
- Older people seem to have a much higher rate of catching stroke (> 67 years)
- Individuals who never smoked surprisingly are most prone to stroke that others
- Obesity, BMI and glucose levels does seem to have some effect
- Patients with heart disease and hypertension are likely to have stroke
- With the data collected, previously married, female, private or self employed, and urban living are some common features that stroke patients have

Data Engineering and Analysis

- BMI column contains 201 null values and hence all these values are replaced with the mean of the BMI column.
- Additionally, the BMI column is feature engineered into having 4 categories, particularly: slim, fit, overweight, obese.
- Similarly, the glucose level column is split to either low, normal or diabetic levels which will provide better information.
- Lastly, for all the categorical columns, the get dummies function is performed whilst dropping the irrelevant columns.



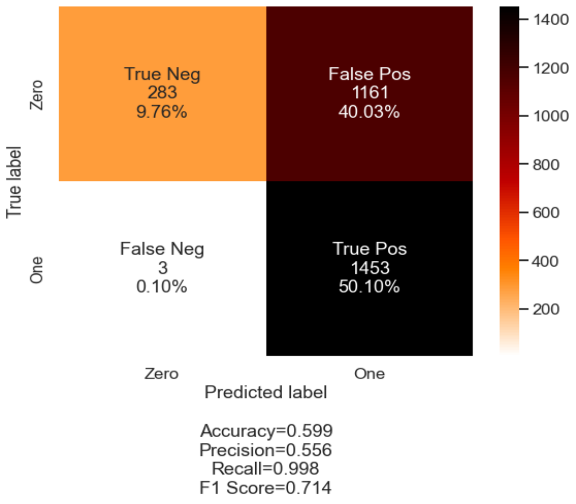


Machine Learning Models

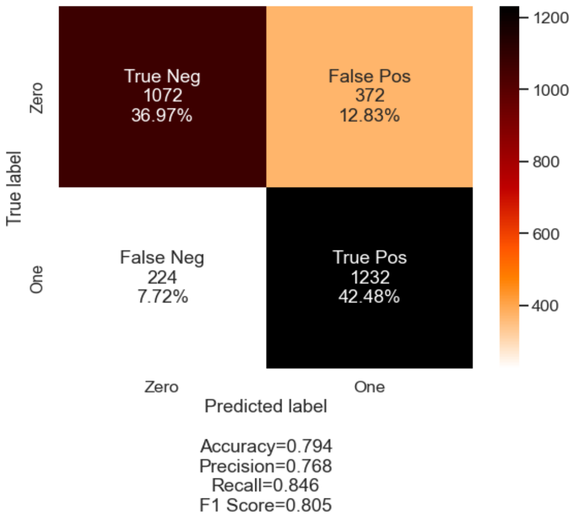
Implementing 9 different machine learning models, we predict (train & test) on the oversampled ADASYN dataset with Naïve Bayes, SGD and SVM performing the best for recall

	Model	Test Accuracy	Metrics
0	Logistic regression	0.841379	precision = 0.86, recall = 0.82, F1 = 0.84
1	Naive Bayes	0.598621	precision = 0.56, recall = 1.00, F1 = 0.71
2	Stochastic Grad Descent (SGD)	0.794483	precision = 0.77, recall = 0.85, F1 = 0.81
3	Random Forest Classifier	0.810345	precision = 0.89, recall = 0.71, F1 = 0.79
4	Gradient Boosting Classifier	0.834483	precision = 0.88, recall = 0.77, F1 = 0.82
5	Support Vector Machine	0.763793	precision = 0.73, recall = 0.84, F1 = 0.78
6	K Nearest Classifier	0.768276	precision = 0.83, recall = 0.68, F1 = 0.75
7	Decison Tree	0.685862	precision = 0.89, recall = 0.43, F1 = 0.58
8	XGBoost	0.794138	precision = 0.84, recall = 0.73, F1 = 0.78

Naïve Bayes



Stochastic Gradient Descent

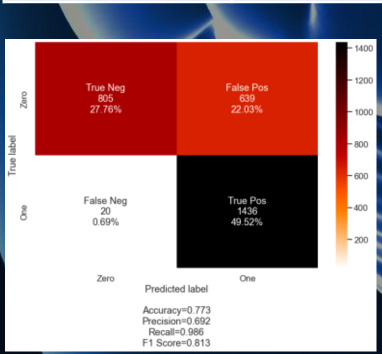


Hyperparameter Tuning & Stacking

After performing Grid Search for various hyperparameters concerning the Naïve Bayes and Stochastic Gradient Descent models, below are the results obtained where Naïve Bayes is the clear winner

Naïve Bayes

	Precision	Recall	F1-score
0	0.98	0.56	0.71
1	0.69	0.99	0.81
accuracy			0.77
Macro avg	0.83	0.77	0.76
Weighted avg	0.83	0.77	0.76

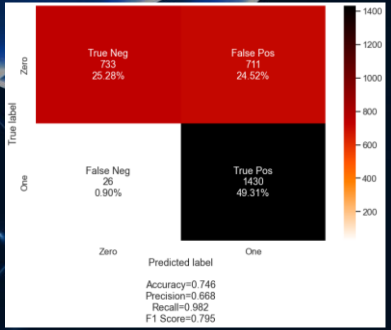


GaussianNB

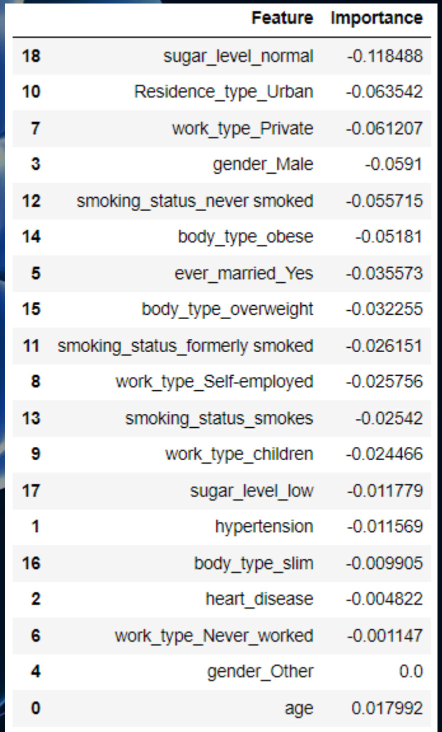
(var_smoothing=6.108761607505 04e-05)

Stochastic Gradient Descent

	Precision	Recall	F1-score
0	0.97	0.51	0.67
1	0.67	0.98	0.80
accuracy			0.75
Macro avg	0.82	0.74	0.73
Weighted avg	0.82	0.75	0.73



{'alpha': 1, 'loss': 'log', 'penalty': 'l2'}



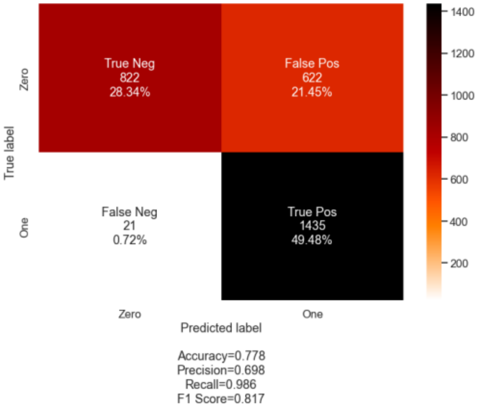
	Feature	Importance
18	sugar_level_normal	-0.118488
10	Residence_type_Urban	-0.063542
7	work_type_Private	-0.061207
3	gender_Male	-0.0591
12	smoking_status_never smoked	-0.055715
14	body_type_obese	-0.05181
5	ever_married_Yes	-0.035573
15	body_type_overweight	-0.032255
11	smoking_status_formerly smoked	-0.026151
8	work_type_Self-employed	-0.025756
13	smoking_status_smokes	-0.02542
9	work_type_children	-0.024466
17	sugar_level_low	-0.011779
1	hypertension	-0.011569
16	body_type_slim	-0.009905
2	heart_disease	-0.004822
6	work_type_Never_worked	-0.001147
4	gender_Other	0.0
0	age	0.017992

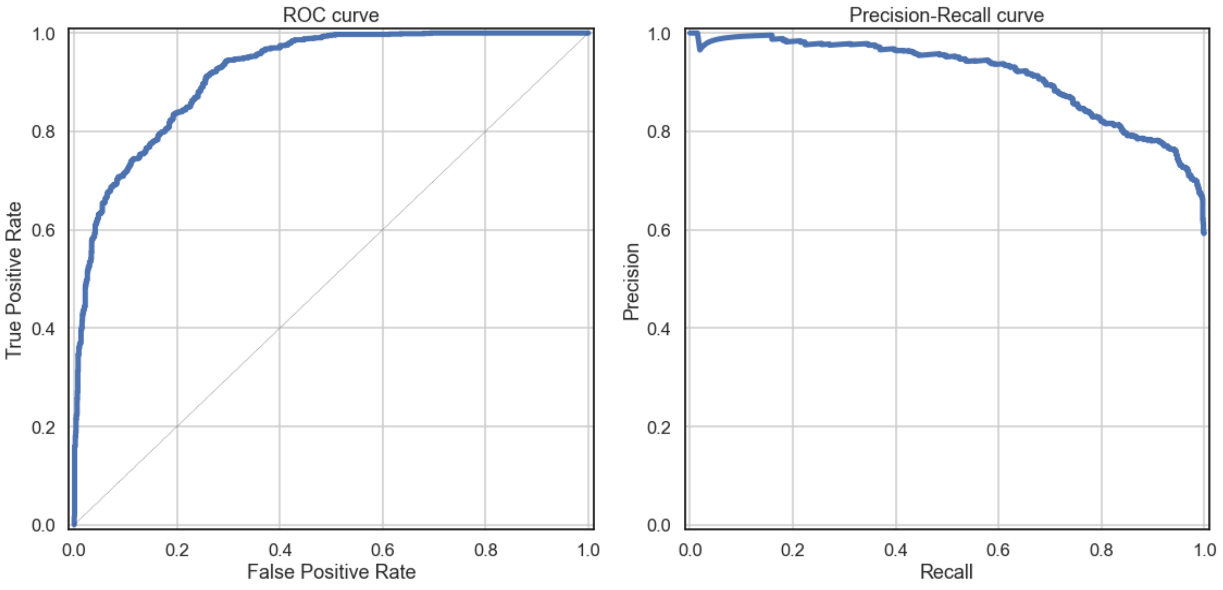
Hyperparameter Tuning & Stacking

After implementing **Naïve Bayes** and **SGD**, Grid Search was performed on a **Gradient Boosting Classifier** and then all these three models were used in a voting classifier with soft voting and weights = (0.1, 0.9, 0.1) where the order is SGD, Naïve Bayes and Gradient Boosting. The results show that the ensemble classifier performed even better than overall (same in recall for minority class) and hence, this model should be selected by hospital stakeholders as it minimizes both false positives and false negatives better than all other models. Below are the classification report, confusion matrix, ROC and precision-recall curve for this stacked model where the balance between precision and recall is more than sufficient here

Voting Classifier (Ensemble Modelling)

	Precision	Recall	F1-score
0	0.98	0.57	0.72
1	0.70	0.99	0.82
accuracy			0.78
Macro avg	0.84	0.78	0.77
Weighted avg	0.84	0.78	0.77



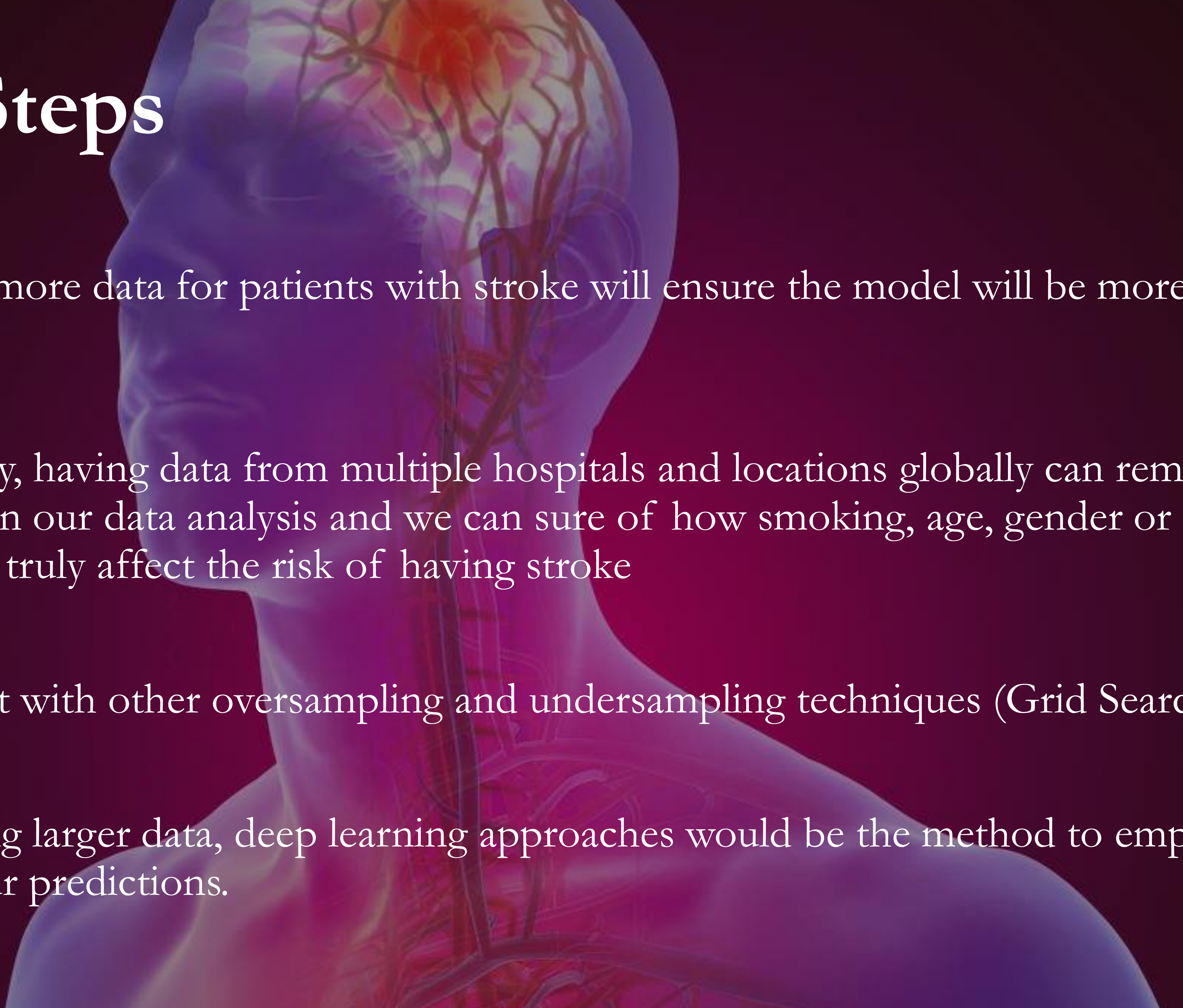


```
{'sgd': SGDClassifier(alpha=1, loss='log', penalty='l1'),

'nb': GaussianNB(var_smoothing=6.108761607505 04e-05),

'gbc': GradientBoostingClassifier(max_depth=5, max_features=7, min_samples_leaf=5, min_samples_split=3, subsample=0.8)}
```

Next Steps



- Collecting more data for patients with stroke will ensure the model will be more reliable
- Additionally, having data from multiple hospitals and locations globally can remove any biases in our data analysis and we can sure of how smoking, age, gender or BMI differences truly affect the risk of having stroke
- Experiment with other oversampling and undersampling techniques (Grid Search)
- After having larger data, deep learning approaches would be the method to employ to enhance our predictions.

Thank You!