Stroke Prediction Analysis

Group 7:

AARON TANG JUN JIE

AHMADUL MATIN BIN AHMAD KAMAL

HO XIN ZHEN

IGNATIUS TANG HONG-QUAN

Table of Contents

1 **Data Preparation**

(AHMADUL MATIN)

2 **Features Selection**

(AARON TANG)

3 **Train & Test Models**

(HO XIN ZHEN)

4 **Model Evaluation & Real-Time Prediction**

(IGNATIUS TANG)

Stroke prediction amongst adults over 30 years

- There are many different factors affecting the chances of getting stroke
- Aim is to predict the possibility of getting a stroke for adults over 30

Chosen Dataset:

Stroke Prediction Dataset

By Fedesoriano

PROBLEM 1:

What defines an at-risk patient?

Exploratory Data Analysis

1. Convert smoking status data to categorical data
2. Check for duplicates in dataset
3. Check for missing values in dataset
4. Drop unusual category in dataset (smoking status)
5. Fill up missing values using mean values

Type of ML Project: Classification

Features Selection

Our team decided to narrow down, using only age, average glucose level, BMI and smoking status.

	age	avg_glucose_level	bmi	smoking_status	stroke
0	67.0	228.69	36.6	formerly smoked	1
1	61.0	202.21	NaN	never smoked	1
2	80.0	105.92	32.5	never smoked	1
3	49.0	171.23	34.4	smokes	1
4	79.0	174.12	24.0	never smoked	1
...	...	...	...	...	...
5105	80.0	83.75	NaN	never smoked	0
5106	81.0	125.20	40.0	never smoked	0
5107	35.0	82.99	30.6	never smoked	0
5108	51.0	166.29	25.6	formerly smoked	0
5109	44.0	85.28	26.2	Unknown	0

Cleaning the Dataset

Our team decided to narrow down, using only age, average glucose level, BMI and smoking status.

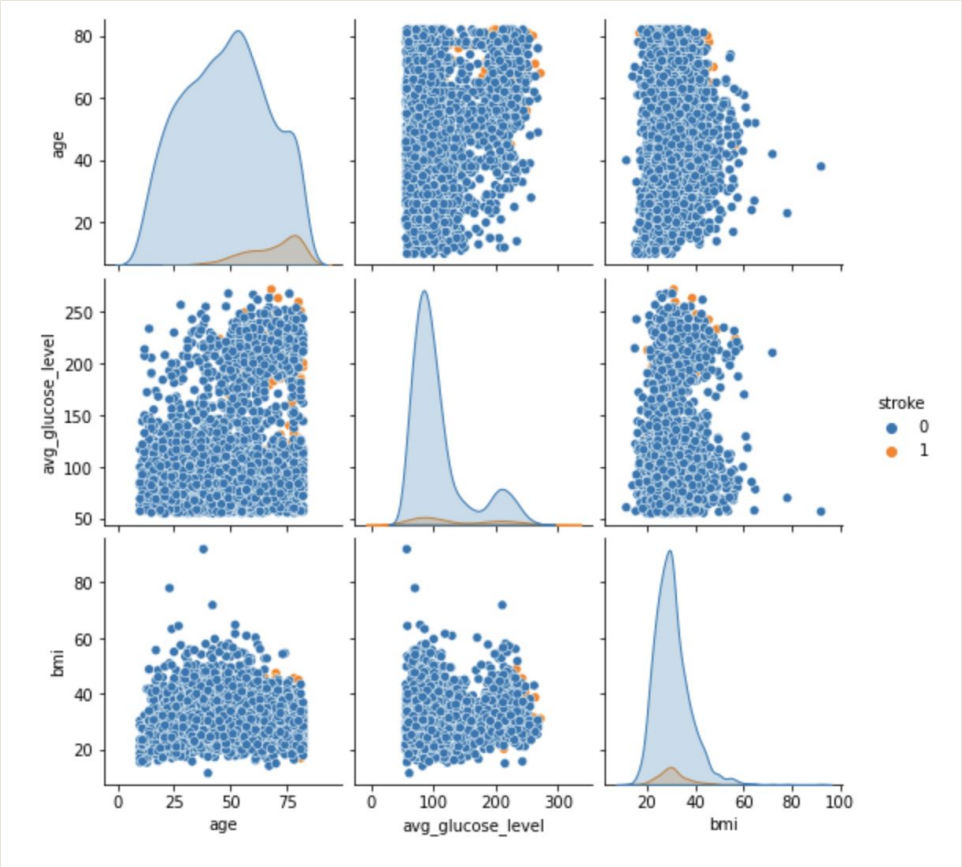
'Total number of of Duplicates present in data: 0'

Number of Missing Values in our data set

	Variable	Missing Values
0	age	0
1	avg_glucose_level	0
2	bmi	125
3	smoking_status	0
4	stroke	0

We will use `mean` to fill up missing values in Numerical Continuous columns(`bmi`)

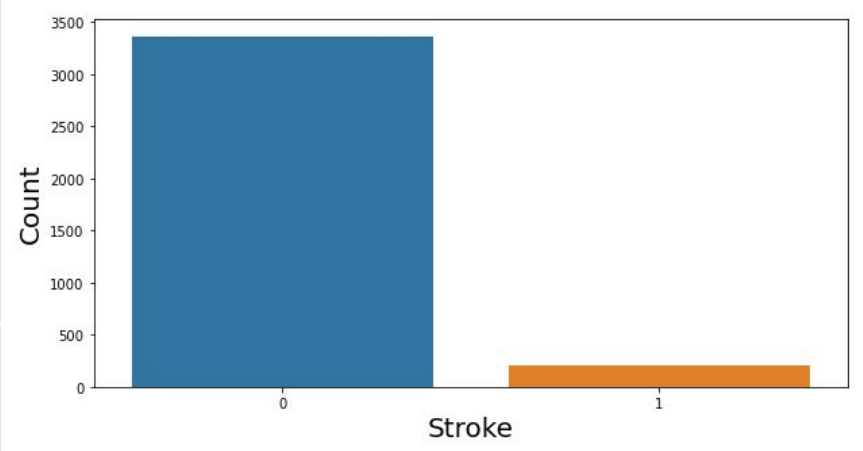
Visualization of variables



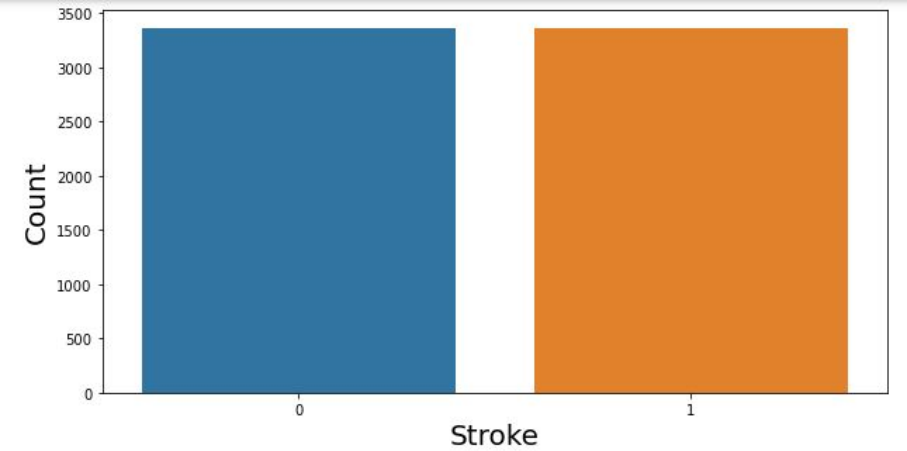
Visualization of variables

Unbalanced stroke data will cause accuracy to be low, hence oversampling was used to balance the class distribution

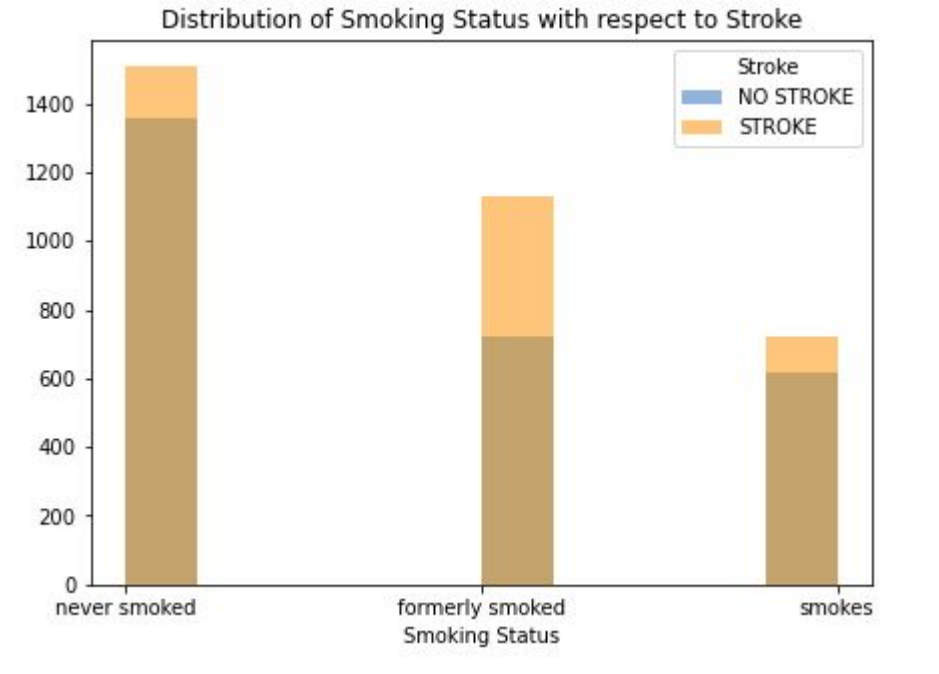
Before oversampling

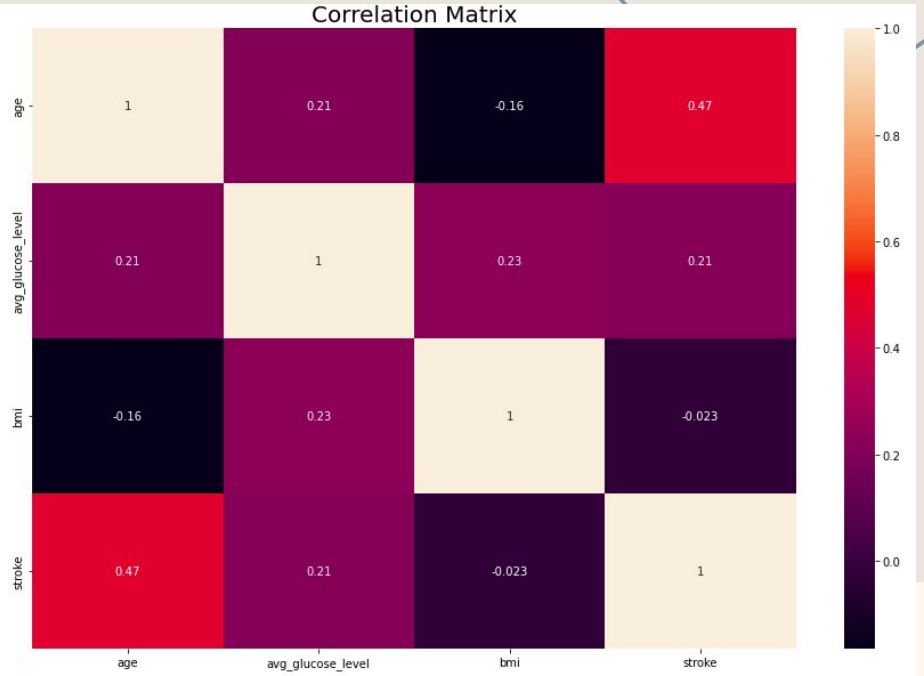


After oversampling



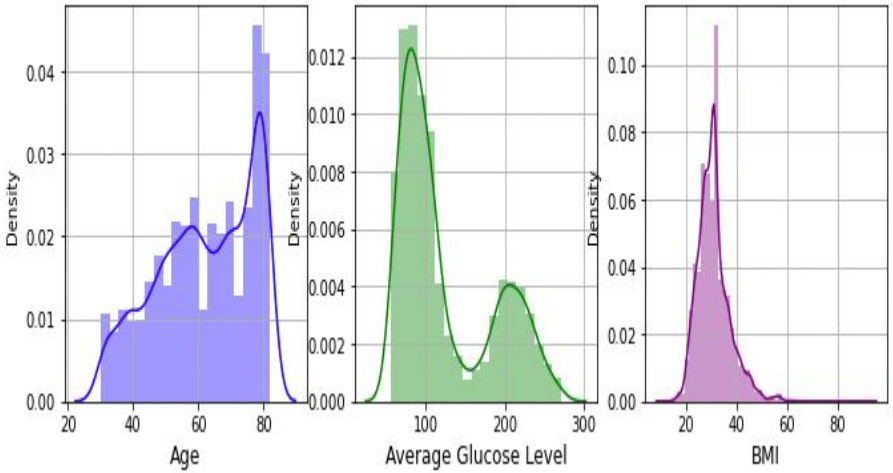
Visualization of variables VS stroke



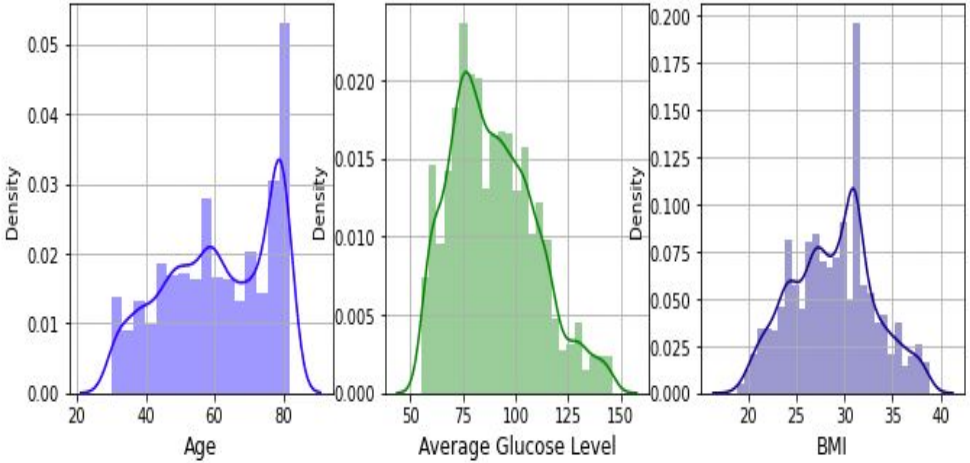


Outlier Analysis and Removal

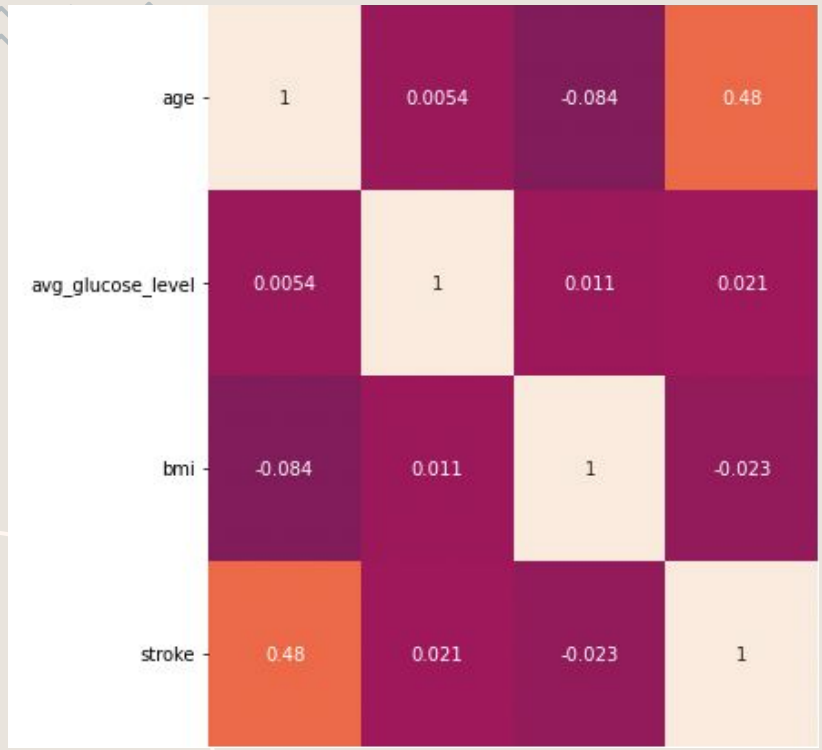
BEFORE

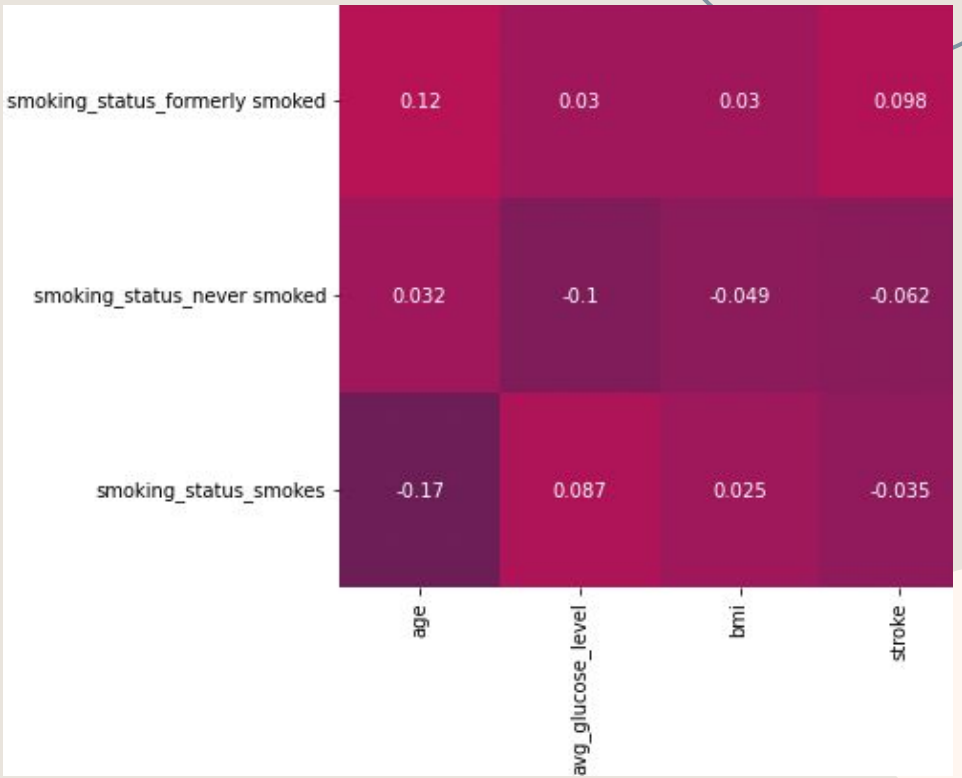


AFTER



Correlation of features after label encoding





PROBLEM 2:

What is the probability of someone above the age of 30 to get a stroke, based on bmi and glucose level?

Classification Model Selection

Logistic Regression:

Predicts based on the *probabilities* of data points on a binary scale.

KNeighbors:

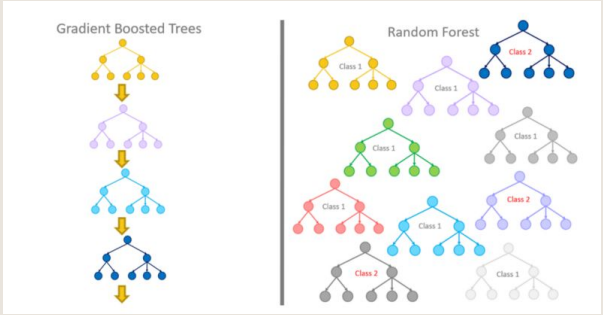
Predicts based on *distance* between train and test points

Decision Tree:

Breaks down dataset into smaller subsets

Gradient Boosting Classifier:

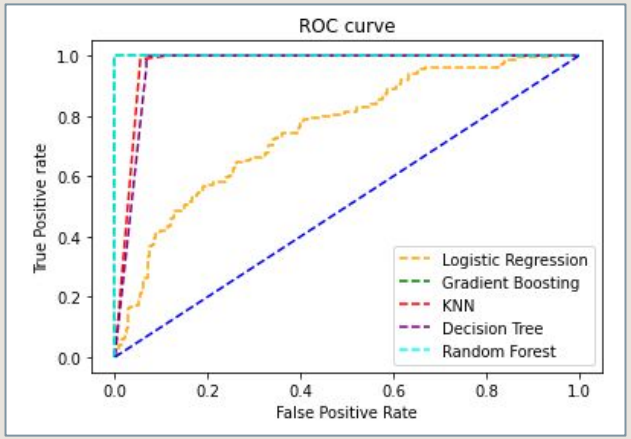
Combines Decision Trees additively



Random Forest Classifier:

Combines multiple Decision Trees

Gradient Boosting Model Evaluation



	Model	Accuracy
0	Logistic Regression	0.673028
1	Gradient Boosting	0.982188
2	KNeighbors	0.923664
3	Decision Tree	0.965649
4	Random Forest	0.979644

Gradient Boosting Model has the highest accuracy score, which is also depicted in the ROC curve where the graph has the largest area-under-graph value. (*very close to Random Forest*)

Gradient Boosting Classifier: Algorithm

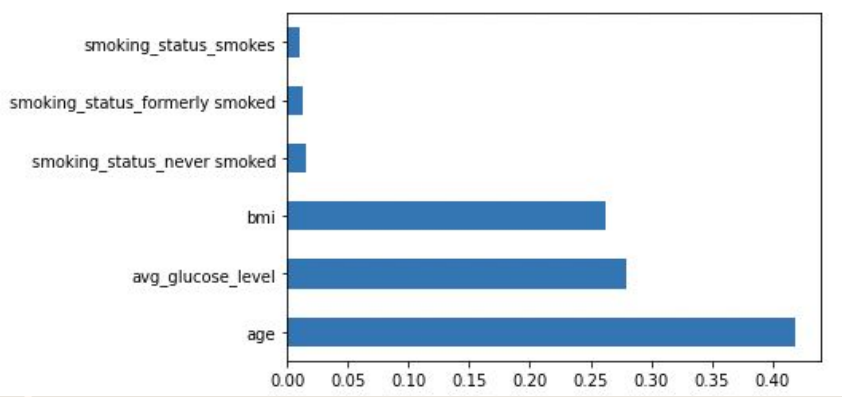
1. Initialize with a predicted value: log (has stroke/ no stroke)
   a. can convert to a probability: $P(\text{has stroke}) = e^{\log(\text{odds})}/(1+e^{\log(\text{odds})})$
2. For m=1 to 100
   a. Find residuals: difference between observed and predicted values
   b. Train regression tree (max depth = 10)
   c. Calculate final output value of leaf with equation below

$$\frac{\sum Residual}{\sum[PreviousProb * (1 - PreviousProb)]}$$

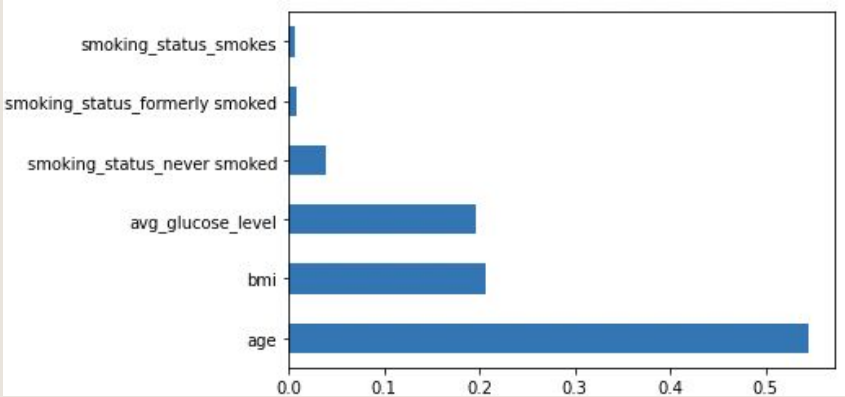
   d. New prediction value = prediction value + (learning rate * output)
   e. Repeat until m = 100

Notable Variables

Age, average glucose level and bmi seems to be the most important across different models.



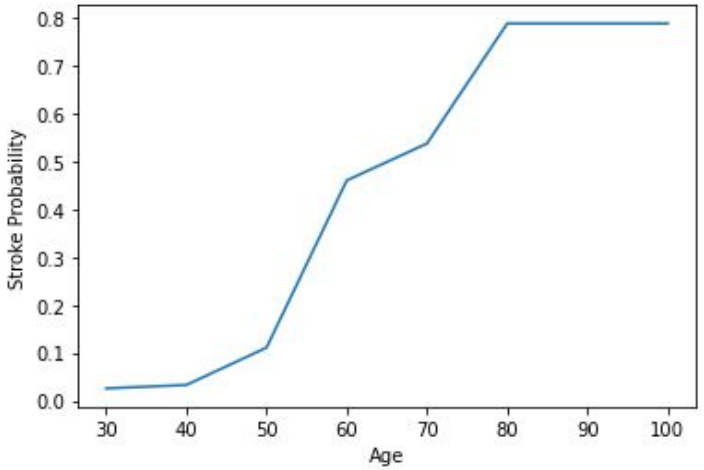
Random Forest Model



Gradient Boosting Classifier

Real-Time Prediction

Low-risk: 0-10%
Medium-risk: 10-15%
High-risk: <15%



Input: [age, avg_glucose_level, bmi]

```
qst = GradientBoostingClassifier(random_state=0).fit(X_train, y_train)
#array input of [age, avg_glucose_level, bmi]
prediction = qst.predict_proba(np.array([[60,112,25]]))[:,1]
print("Probability of person getting a stroke is: ", prediction)
print( "Probability of person not getting a stroke is: ", 1-prediction)
```

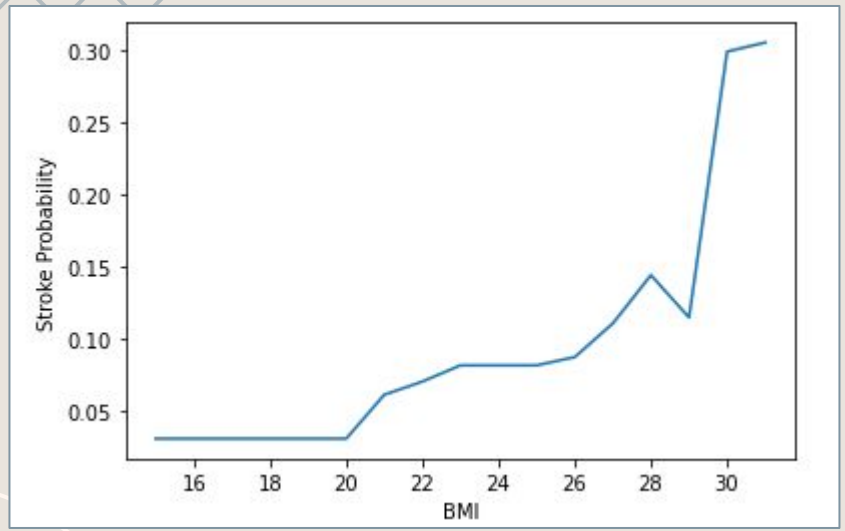
```
Probability of person getting a stroke is:  [0.46000482]
Probability of person not getting a stroke is:  [0.53999518]
```

```
if prediction < 0.10:
    print("Person is at low risk of getting a stroke within next 5 years.")
if 0.10 < prediction < 0.15:
    print("Person is at medium risk of getting a stroke within next 5 years.")
if 0.15 < prediction:
    print("Person is at high risk of getting a stroke within next 5 years.")
```

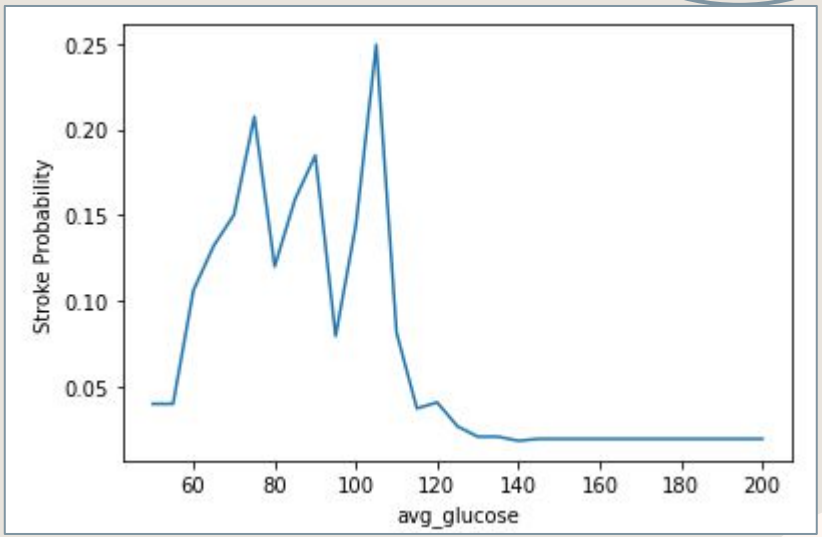
```
Person is at high risk of getting a stroke within next 5 years.
```

From the age-stroke_probability graph above, the rate of increase in stroke probability is the highest at age 50-60.
Patients becomes at mild-risk of stroke at age 45, and high-risk at 55. From this, we can conclude that the critical period to be check and tested for stroke symptoms is the age of 45 where the rate of increase in stroke probability is still relatively manageable.

Interesting Features



To maintain a low-risk stroke probability at 45, one must keep their **BMI** below 26.



The **avg_glucose** to stroke probability graph is haphazard, no distinct trend or pattern shown to draw any reliable conclusions.

Conclusion

Problem 1: What defines an at-risk patient?

Stroke Probability risk categories:

- **Low-risk: 0-10%**
- **Medium-risk: 10-15%**
- **High-risk: >15%**

Problem 2: What is the probability of stroke of a person above the age of 30?

(keeping bmi and avg_glucose to mean)

- **Low-risk: age 30 - 45**
- **Medium-risk: age 45 - 50**
- **High-risk: age 50 and above**

References

1. Kurama, V. (2021, April 9). *Gradient boosting for classification*. Paperspace Blog. Retrieved November 5, 2022, from https://blog.paperspace.com/gradient-boosting-for-classification/
2. Department of Health & Human Services. (2013, April 29). *Heart disease and stroke - your risk score*. Better Health Channel. Retrieved November 5, 2022, from https://www.betterhealth.vic.gov.au/health/conditionsandtreatments/heart-disease-and-stroke-your-risk-score
3. Kogan, E., Twyman, K., Heap, J., Milentijevic, D., Lin, J. H., & Alberts, M. (2020). Assessing stroke severity using electronic health record data: A machine learning approach. *BMC Medical Informatics and Decision Making*, *20*(1). https://doi.org/10.1186/s12911-019-1010-x
4. *Understanding random forest - towardsdatascience.com*. (n.d.). Retrieved November 5, 2022, from https://towardsdatascience.com/understanding-random-forest-58381e0602d2
5. Naukri.com. (n.d.). Retrieved November 5, 2022, from https://www.naukri.com/learning/articles/one-hot-encoding-vs-label-encoding/#Label-encoding
6. Soriano, Fede. “Stroke Prediction Dataset.” *Kaggle*, 26 Jan. 2021, https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset.