Table of Contents

[Overview 1](#_Toc98948934)

[Potential Uses 2](#_Toc98948935)

[Project Description 2](#_Toc98948936)

[Data Collection 3](#_Toc98948937)

[Assam Health Survey Dataset 3](#_Toc98948938)

[Heart Risk Dataset 7](#_Toc98948939)

[Stroke Risk Dataset 9](#_Toc98948940)

[Data Processing 10](#_Toc98948941)

[Heart Risk Dataset 10](#_Toc98948942)

[Exploration and Cleaning 10](#_Toc98948943)

[Testing 14](#_Toc98948944)

[Modelling 15](#_Toc98948945)

[Stroke Risk Dataset 17](#_Toc98948946)

[Exploration and Cleaning 17](#_Toc98948947)

[Modelling 18](#_Toc98948948)

[Assam Health Survey Dataset 20](#_Toc98948949)

[Data Cleaning 21](#_Toc98948950)

[Classification 22](#_Toc98948951)

[Data Exploration & Statistical Testing 22](#_Toc98948952)

[Exporting of Classification 30](#_Toc98948953)

# Overview

* Built a classifier model and detected all residents by their ID from all districts of Assam who might be at a higher risk of stroke.
* Downloaded over 140,000+ observations of Health Survey Data with 53 unique features comprising from 23 districts of Assam from <https://data.gov.in/> to perform classification on.
* Cleaned, explored and manipulated the entire dataset extensively on Python to make it usable for our model.
* Engineered new features and performed deep exploration on the data.
* Used 2 different datasets from separate sources to train the stroke risk classifier model.
* Applied Synthetic Minority Oversampling Technique (SMOTE) and various other methods to overcome extreme imbalance ( 1.8% Minority Proportion ) in the stroke risk dataset before modelling the classifier.
* Built a model that was able to detect the minority class with 75% accuracy.

## Potential Uses

* People classified under risk category can be reached out to spread awareness, provide diagnostic tools, medical help, etc. to help them take care of their health.
* The information can be used by insurance companies, health diagnostic companies, pharma companies, hospitals, etc. to directly reach out and find potential clients who might buy their products and services.

# Project Description

A stroke is a serious life-threatening medical condition that happens when the blood supply to part of the brain is cut off. Strokes are a medical emergency and urgent treatment is essential. But if a person is made aware of the risk and starts taking precautions and makes lifestyle changes, strokes can be prevented significantly.

According to CDC, USA ( [Source](https://www.cdc.gov/stroke/facts.htm#:~:text=Every%20year%2C%20more%20than%20795%2C000,are%20first%20or%20new%20strokes.&text=About%20185%2C000%20strokes%E2%80%94nearly%201,have%20had%20a%20previous%20stroke.&text=About%2087%25%20of%20all%20strokes,to%20the%20brain%20is%20blocked.) ),

Every year, more than 795,000 people in the United States have a stroke. About 610,000 of these are first or new strokes. About 185,000 strokes—nearly 1 of 4—are in people who have had a previous stroke.

In this project we will be building a classifier model to detect residents from 23 districts of Assam who might be at a higher risk of stroke.

This can be very beneficial for the residents as they can be more self-aware and start taking preventative measures and make lifestyle changes.

On the other hand, this can also work as a very crucial piece of information for insurance companies, health diagnostic companies, pharma companies, hospitals, etc. to directly reach out to and maybe find potential clients.

We will be using multiple datasets of each individual districts downloaded from <https://data.gov.in> and concatenating them together into one dataset for our use case. We will also be using 2 separate datasets, heart disease risk data and stroke risk data from separate sources to build our Stroke Risk Classifier model.

**It would also be interesting to explore other features of the Assam Health Survey dataset for both the stroke risk and non-risk groups and see if there are any differences or points of interest among them.**

# Data Collection

## Assam Health Survey Dataset

All the Assam Health Survey Data are sourced from <https://data.gov.in> which is the Open Government Data (OGD) Platform of India. The datasets were downloaded with consent from the website which was provided through a form.

We received 23 datasets for the following 23 districts of Assam:

1. Jorhat

2. Nagaon

3. Golaghat

4. Kamrup

5. Sibsagar

6. Bongaigaon

7. Hailakandi

8. Cachar

9. Barpeta

10. Goalpara

11. Karimganj

12. Tinsukia

13. Dhubri

14. Dhemaji

15. Dibrugarh

16. Karbi Anglong

17. Darrang

18. North Cachar Hills

19. Kokrajhar

20. Lakhimpur

21. Nalbari

22. Marigaon

23. Sonitpur

The datasets came with the following 53 features:

1. state\_code

2. district\_code

3. rural\_urban

4. stratum

5. PSU\_ID

6. ahs\_house\_unit

7. house\_hold\_no

8. date\_survey

9. test\_salt\_iodine

10. record\_code\_iodine

11. record\_code\_iodine\_reason

12. sl\_no

13. Sex

14. usual\_residance

15. usual\_residance\_Reason

16. identification\_code

17. Age\_Code

18. Age

19. date\_of\_birth

20. month\_of\_birth

21. year\_of\_birth

22. Weight\_measured

23. Weight\_in\_kg

24. Length\_height\_measured

25. length\_height\_code

26. Length\_height\_cm

27. Haemoglobin\_test

28. Haemoglobin

29. Haemoglobin\_level

30. BP\_systolic

31. BP\_systolic\_2\_reading

32. BP\_Diastolic

33. BP\_Diastolic\_2reading

34. Pulse\_rate

35. Pulse\_rate\_2\_reading

36. Diabetes\_test

37. fasting\_blood\_glucose

38. fasting\_blood\_glucose\_mg\_dl

39. Marital\_status

40. gauna\_perfor\_not\_perfor

41. duration\_pregnanacy

42. first\_breast\_feeding

43. is\_cur\_breast\_feeding

44. day\_or\_mn\_for\_breast\_feeding\_cd

45. day\_or\_month\_for\_breast\_feeding

46. water\_month

47. ani\_milk\_month

48. semisolid\_month\_or\_day

49. solid\_month

50. vegetables\_month\_or\_day

51. illness\_type

52. illness\_duration

53. treatment\_type

#### Univariate Stats First Glance

After concatenating all the 23 datasets, we get the following Univariate Stats for all features.



## Heart Risk Dataset

The dataset for heart risk classification is created by combining different datasets already available independently but not combined before. In this dataset, 5 heart datasets are combined over 11 common features. The five datasets used for its curation are:

Cleveland: 303 observations

Hungarian: 294 observations

Switzerland: 123 observations

Long Beach VA: 200 observations

Stalog (Heart) Data Set: 270 observations

Total: 1190 observations

Duplicated: 272 observations

**Final dataset: 918 observations**

Every dataset used can be found under the Index of heart disease datasets from UCI Machine Learning Repository on the following link:

<https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/>

The dataset came with the following features:

1. Age

2. Sex

3. ChestPainType

4. RestingBP

5. Cholesterol

6. FastingBS

7. RestingECG

8. MaxHR

9. ExerciseAngina

10. Oldpeak

11. ST\_Slope

12. HeartDisease

#### Univariate Stats First Glance

## Stroke Risk Dataset

The dataset for stroke risk classification is sourced from <https://data.mendeley.com/datasets/x8ygrw87jw/1>

This dataset is extremely imbalanced with a 1 : 55 minority to majority class ratio and comes with a total of 43400 observations. We will be applying various measures to balance the dataset going forward.

The dataset came with the following features:

1. id

2. gender

3. age

4. hypertension

5. heart\_disease

6. ever\_married

7. work\_type

8. Residence\_type

9. avg\_glucose\_level

10. bmi

11. smoking\_status

12. stroke

#### Univariate Stats First Glance



# Data Processing

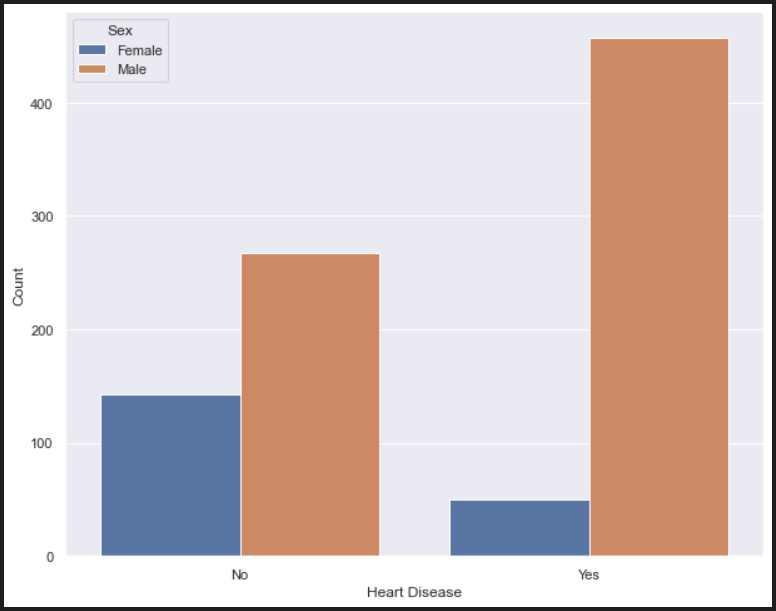
## Heart Risk Dataset

### Exploration and Cleaning

The data cleaning process was fairly simple and straight forward for the Heart risk dataset as this dataset was already pre-processed and cleaned from the get-go.

#### Sex

We plotted the data in the feature on a bar graph to visually explore the information further.



We see that ratio of Males is predominantly higher than Females in our data and Males are significantly more prone to heart disease than Females.

Next, we create dummies for our feature name and proceed ahead.

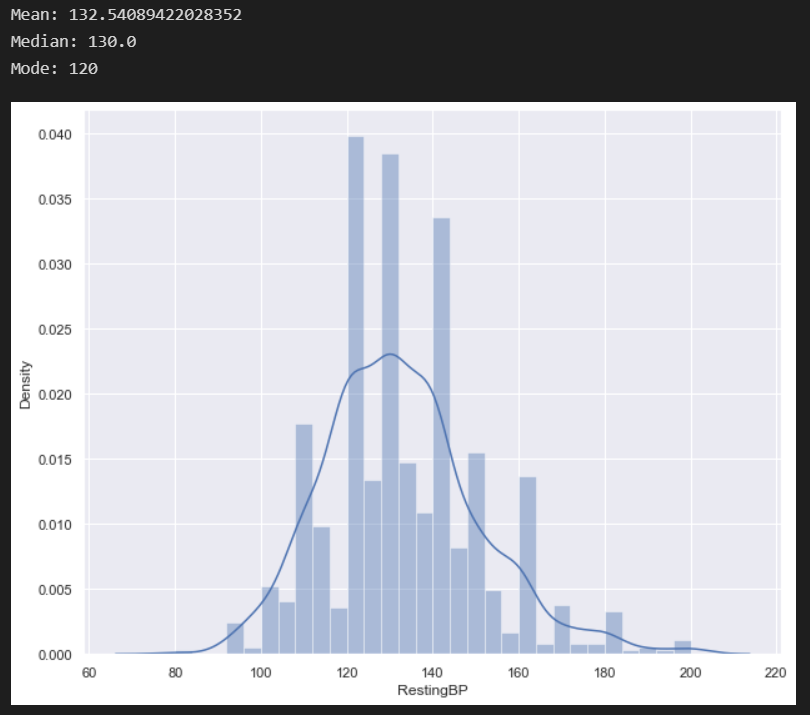
#### RestingBP

We detected a discrepancy with the feature “RestingBP” in the dataset.

According to Centre of Disease Control and Prevention ( [Source](https://www.cdc.gov/bloodpressure/about.htm#:~:text=Blood%20pressure%20is%20measured%20using,your%20heart%20rests%20between%20beats.) ),

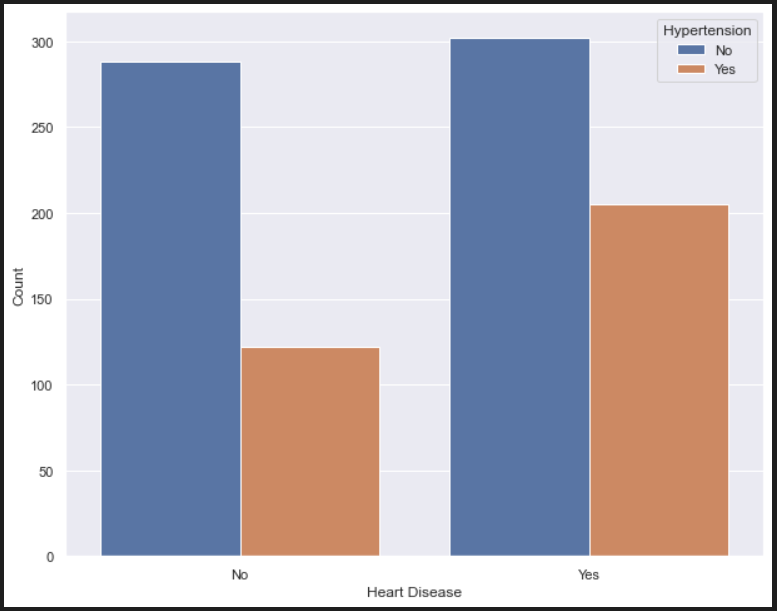
“Blood pressure is measured using two numbers: The first number, called systolic blood pressure, measures the pressure in your arteries when your heart beats. The second number, called diastolic blood pressure, measures the pressure in your arteries when your heart rests between beats.”

But upon exploring the information provided in our dataset by checking its measures of central tendency and distribution plot, we see that we are given the systolic blood pressure and not the diastolic blood pressure as the feature name implies.



So, we proceeded to classify everyone with the assumption that it is systolic blood pressure with anyone above 140 as hypertensive in accordance with the guidelines from the CDC website provided above.

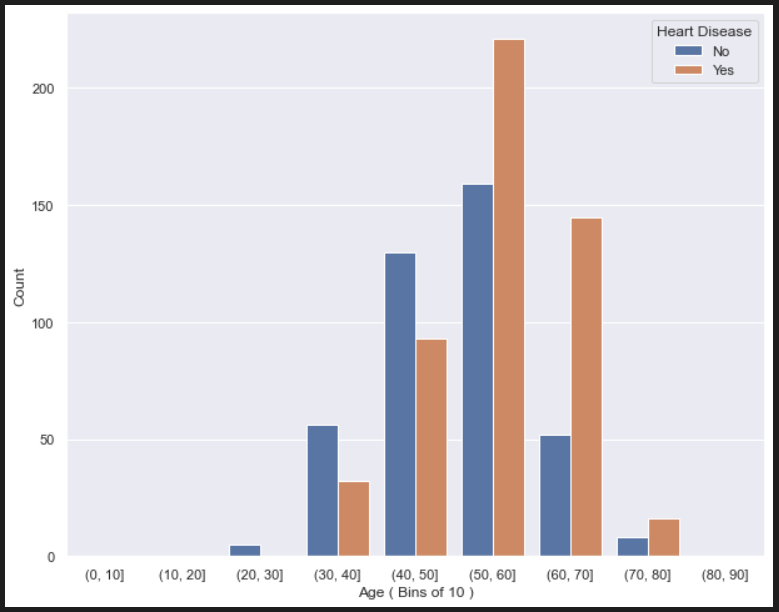
Then, we plot the information on a bar graph to visually explore the information further.



We see that people with hypertension are more prone to having heart disease compared to people without hypertension.

#### Age

We created bins of 10 years and plotted them on a bar graph to visually explore the information.



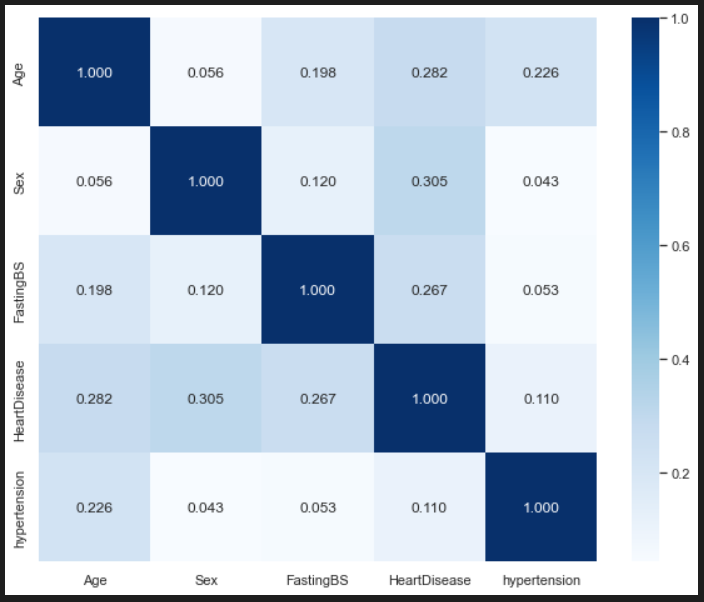
We can see that majority of our data is of people from 30 to 80 years of age and the people of higher age ( 50+ ) are more susceptible to heart disease.

**We keep a note about this going forward as we only want to use classifications in people between 30 to 80 years of age.**

### Testing

#### Pearson Correlation

We check the correlation of our relevant features with each other and plot them on a heatmap.



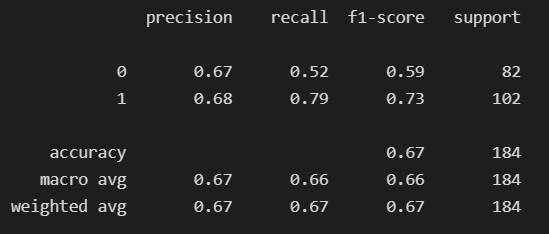
We see most of our features have relatively reliable correlation with our target variable.

### Modelling

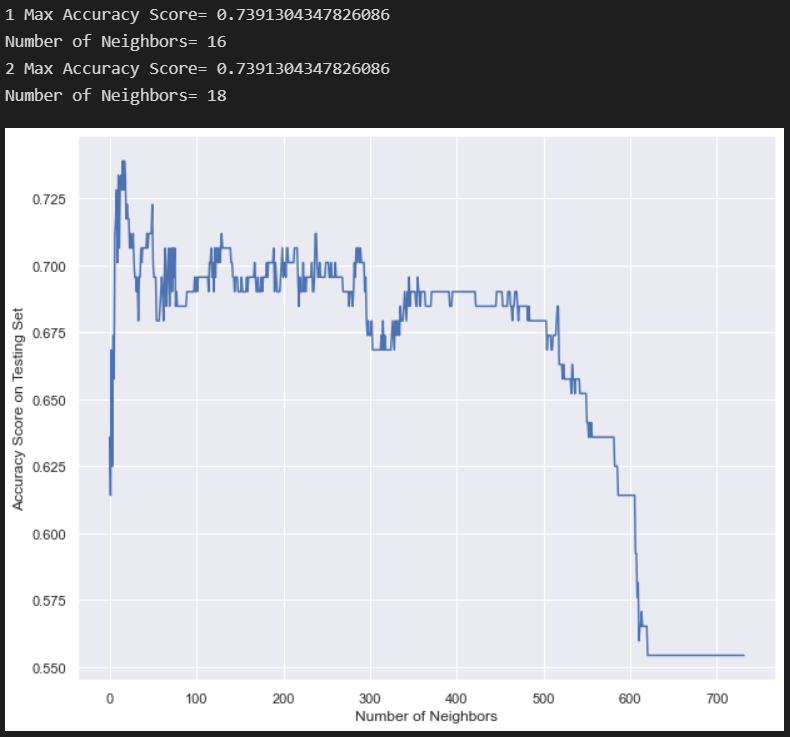
We start our modelling process by scaling our data and splitting it into training and testing sets.

We decided to use k-nearest neighbors (KNN) classifier to build our classifier model because our dataset is small and simple and the classifier is easy to implement and intuitive to understand. Also, because our data is already cleaned of outliers which could cause issues in KNN Classifier, it makes it even more suitable to use KNN to produce accurate results reliably.

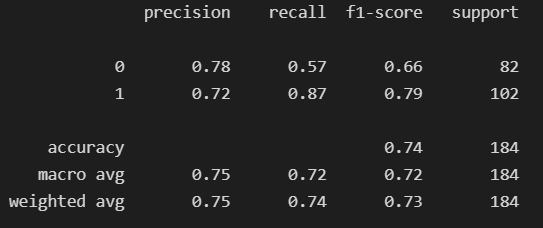
We fit our training data on the default KNN Classifier with 5 nearest neighbors and evaluated our performance as baseline.



Then, we ran a For-Loop to find the optimal value of nearest neighbors that gives the best accuracy scores on our testing set and plotted the results on a line graph along with the number of neighbors which gave the best accuracy score.



We see that 16 and 18 number of neighbors gave the best accuracy score on the testing set. We decided to use either of them ( 18 in our case ) to fit our training data and evaluated the performance on the testing dataset.

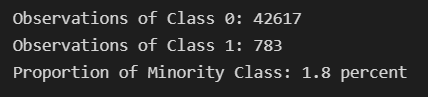


We see that there is a significant improvement in performance compared to baseline especially in detection(recall) and precision scores.

Hence, we decided to save this classifier and proceeded next to our Stroke Risk Dataset.

## Stroke Risk Dataset

This dataset is very interesting in the sense that it is extremely imbalanced with the proportion of minority class being 1.8%. We will be applying various measures to try to balance the dataset before making the classification model.



### Exploration and Cleaning

First, we filter out our dataset with age range of 30 to 80 years only in accordance to our exploration of Age feature in our Heart Risk Dataset.

Then, we start exploring and cleaning as we go through all our relevant features.

#### BMI

We filled the null values in the feature with the mean of BMI of the individual groups of if the person had or not had a stroke.

#### Gender

We dropped "Other" category of gender from our observations as our classifier is not trained for that category of information.

#### High Blood Sugar

We created a new feature which classifies anyone with an average blood sugar level of 140 and above as high blood sugar.

According to [Source](https://www.singlecare.com/blog/normal-blood-glucose-levels/) :

Normal blood sugar range for adults is 100-180 mg/dL so we take the average (140) and classify anyone higher as 1 else 0.

### Modelling

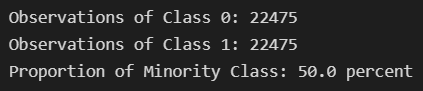
We feature engineered a new feature of high risk of heart disease with our classifier built from Heart Risk Dataset and started our modelling process by scaling our data and splitting it into training and testing sets.

#### SMOTE ( Synthetic Minority Over-Sampling Technique )

With the objective of balancing our Dataset we imported SMOTE from Imbalanced-learn ( imported as imblearn ) library in Python.

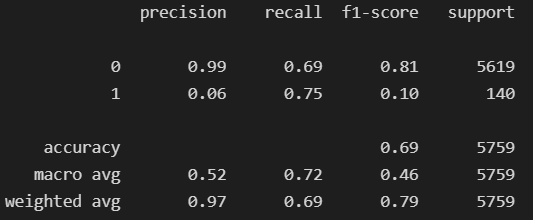
SMOTE is a widely used technique that works by selecting examples that are close in the feature space, drawing a line between the examples in the feature space and drawing a new sample at a point along that line.

After application of SMOTE on our training set, we see that we have a balanced number of observations from each class.



Next, we decided to use Logistic Regression to build our classifier model because our dataset is quite simple and Logistic Regression is easier to implement, interpret, and very efficient to train. Also, since It is very fast at classifying unknown records, it will come in very handy when classifying our Assam Health Survey Dataset which has a higher number of observations.

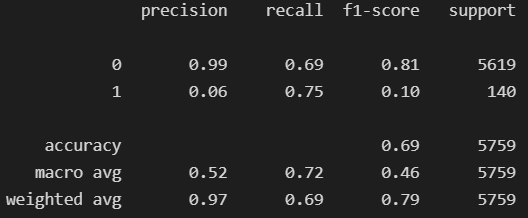
We fit our training data on the default Logistic Regression and evaluated the performance on the testing set.



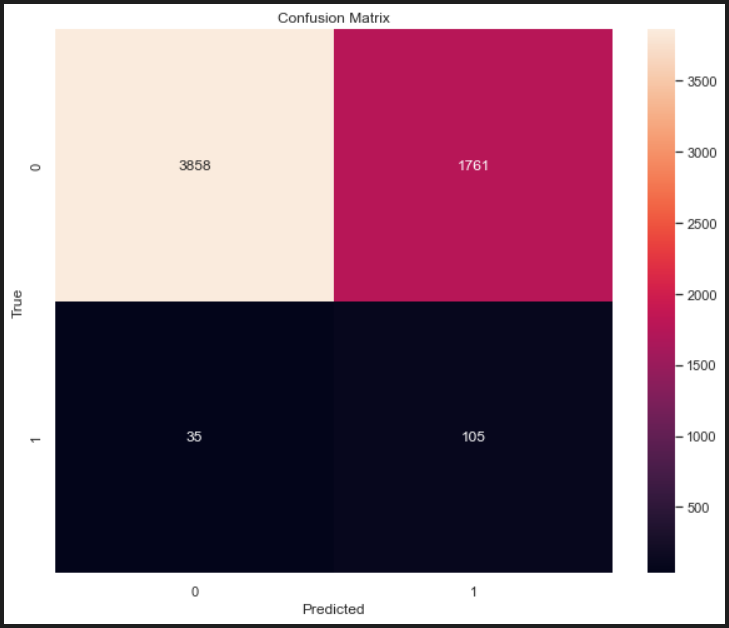
We see that although our precision is low, we have a pretty good score for recall (detection) for our minority class which is what matters most in our case.

In simpler words, we were able to detect 75% of our minority class in an extremely imbalanced dataset. Lower precision means that yes, we also misclassified some data from majority class as minority class but that is something to be dealt ahead later.

We also perform a Grid Search to find the optimal solver and C values for our Logistic Regression. Then we evaluate the performance with the optimal hyper-parameters.



We see that there is no major improvement in performance. We plot a Confusion Matrix to evaluate the classification visually.



We see that out of 140 observations of minority class, we were able to detect 105 of them successfully. We found the performance satisfactory and thus decided to save it to use when classifying our Assam Health Survey dataset.

## Assam Health Survey Dataset

This dataset is very dirty from the get-go as expected since it was downloaded from the official government website of open data.

It has a total of 142838 observations from 23 districts of Assam and has a lot of missing, corrupt and irregular information in the dataset that we will need to look out for while working.

### Data Cleaning

First, we filter out our dataset with age range of 30 to 80 years only in accordance to what our classifiers are trained on. We also filter out the observations that do not have the crucial information our classifier will need to classify.

Then, we start exploring and cleaning as we go through all our relevant features.

#### BP Systolic Reading

In our dataset we have two readings for systolic blood pressure. We take the average of both the readings into the main reading or only the first reading if the second reading is null.

#### Hypertension

We created a feature hypertension where we classified anyone with BP systolic reading of 140 and above as hypertensive.

#### Marital Status

We filled the null values in Marital status with “Never married”. Then, we created a new feature and classified on the basis if a person ever married or not.

#### BMI

We engineered a new feature BMI with the information provided in Height and Weight after removing the outliers.

#### High Blood Sugar

We classify anyone having glucose level of 120 and above as having High Blood Sugar in accordance with our stroke risk classifier model.

#### Heart Disease

We create a new feature which classifies if the person has higher risk of heart disease with our heart risk classifier model.

### Classification

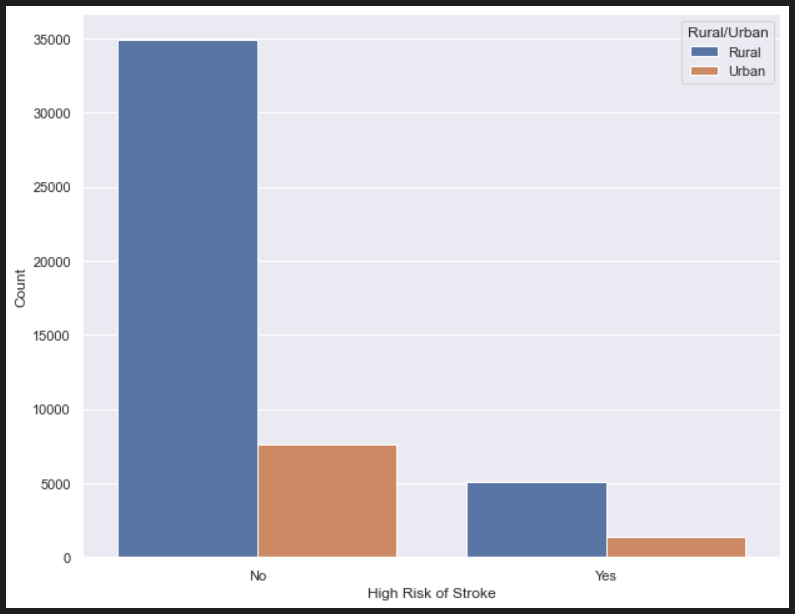
We use our stroke risk classifier model for classification.

With it we classify everyone with a stroke probability of 60% or more as a high risk of stroke individual and store that information back into dataset for further visual explorations.

### Data Exploration & Statistical Testing

#### Rural-Urban Relationship with Stroke Risk Group

We plot the rural or urban feature on the basis of higher risk of stroke group on a bar graph and try to visually explore the information.



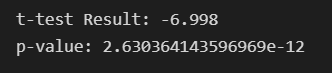
We see that the ratio of observations from rural areas is predominantly higher than urban in our dataset.

But it is not easy to understand if being from rural or urban has any impact on having higher risk of stroke. Hence, to test this out we do a hypothesis test:

**Null Hypothesis (H0):** Being from Rural or Urban has no significant impact on having difference in risk of stroke.

**Alternate Hypothesis (HA):** Being from Rural or Urban has a significant impact on having difference in risk of stroke.

We get the following result after running a t-test:



Since, p-value < 0.05 (alpha), we reject the Null Hypothesis (H0) and conclude that being from rural or urban has a statistically significant effect on having difference in risk of stroke.

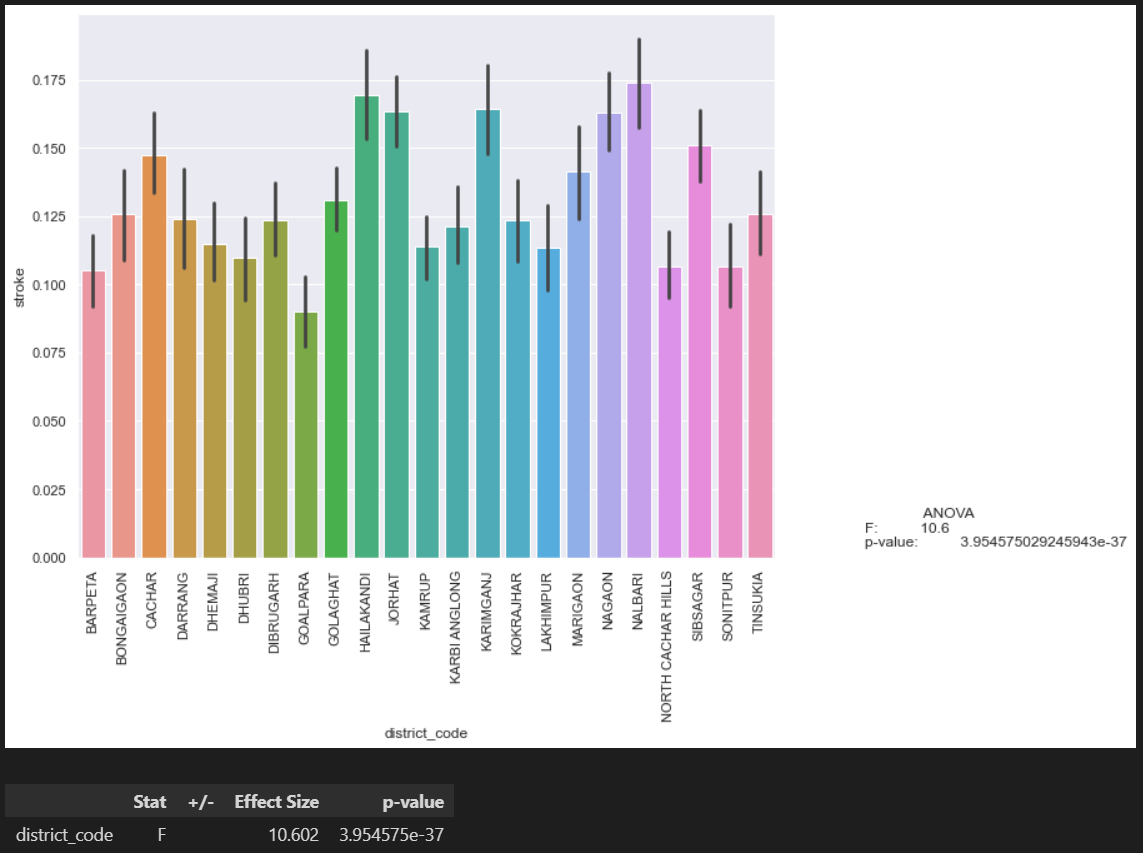
#### Districts Relationship with Stroke Risk Group

To understand if being from different districts has any statistically significant impact on having higher risk of stroke, we perform a hypothesis test.

**Null Hypothesis (H0):** Being from different districts has no significant impact on having higher risk of stroke.

**Alternate Hypothesis (HA):** Being from different districts has significant impact on having higher risk of stroke.

We get the following result after performing One-Way ANOVA test:

**

Since, p-value < 0.05 (alpha), we reject the Null Hypothesis (H0) and conclude that being from different districts has a statistically significant effect on having difference in risk of stroke.

We also perform a Tukey's HSD (honestly significant difference) test between each pair of districts to find the pairs that have the most statistically significant effect on having higher risk of stroke.

Top 10 most significant pairs are as follows:



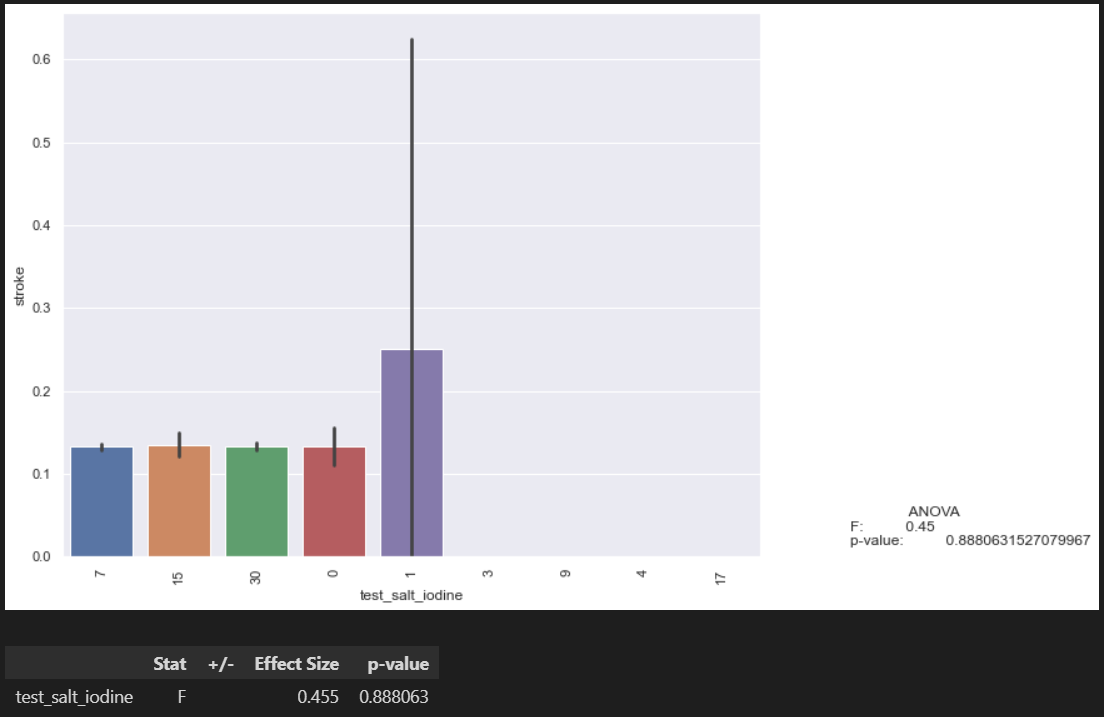
#### Salt Iodine Test Values Relationship with Stroke Risk Group

To understand if having different salt iodine test values has any statistically significant impact on having higher risk of stroke, we perform a hypothesis test.

**Null Hypothesis (H0):** Having different salt iodine test values has no significant impact on having higher risk of stroke.

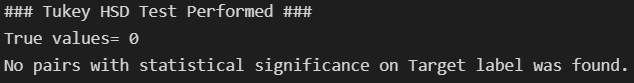
**Alternate Hypothesis (HA):** Having different salt iodine test values has a significant impact on having higher risk of stroke.

We get the following result after performing One-Way ANOVA test:



Since, p-value > 0.05 (alpha), we fail to reject the Null Hypothesis (H0) and conclude that having different salt iodine test values has no statistically significant effect on having difference in risk of stroke.

We also perform a Tukey's HSD (honestly significant difference) test between each pair to verify our finding and we see that indeed there is no statistical significance.



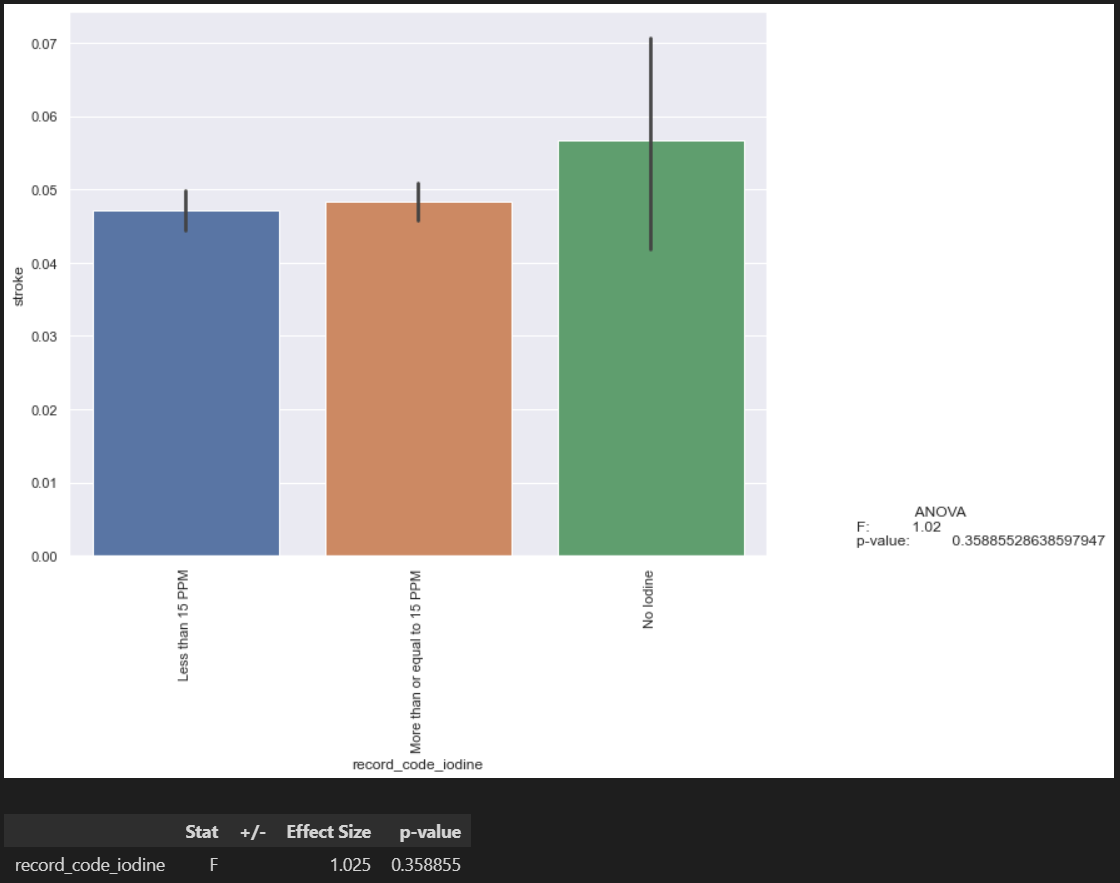
#### Record Code Iodine Values Relationship with Stroke Risk Group

To understand if having different record code iodine values has any statistically significant impact on having higher risk of stroke, we perform a hypothesis test.

**Null Hypothesis (H0):** Having different record code iodine values has no significant impact on having higher risk of stroke.

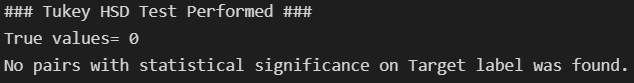
**Alternate Hypothesis (HA):** Having different record code iodine values has a significant impact on having higher risk of stroke.

We get the following result after performing One-Way ANOVA test:



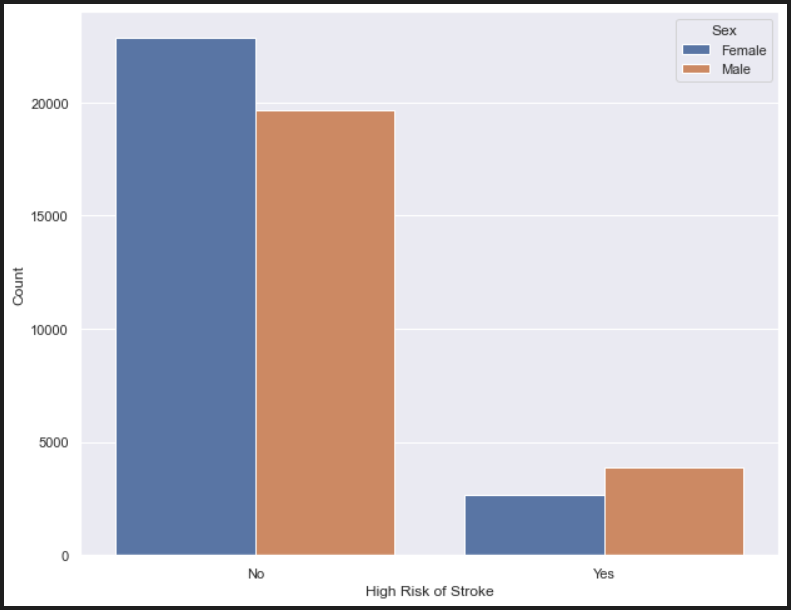
Since, p-value > 0.05 (alpha), we fail to reject the Null Hypothesis (H0) and conclude that having different salt iodine test values has no statistically significant effect on having difference in risk of stroke.

We also perform a Tukey's HSD (honestly significant difference) test between each pair to verify our finding and we see that indeed there is no statistical significance.



#### Sex Relationship with Stroke Risk Group

We plot the sex feature on the basis of higher risk of stroke group on a bar graph and try to visually explore the information.



We see that the ratio of observations of male and female are pretty balanced in our dataset.

We can also see that males are more likely to be in stroke risk category compared to females. To further verify this, we do a hypothesis test:

**Null Hypothesis (H0):** Being Male or Female has no significant impact on having difference in risk of stroke.

**Alternate Hypothesis (HA):** Being Male or Female has a significant impact on having difference in risk of stroke.

We get the following result after running a t-test:



Since, p-value < 0.05 (alpha), we reject the Null Hypothesis (H0) and conclude that being male or female has a statistically significant effect on having a difference in risk of stroke.

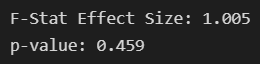
#### Haemoglobin Level Relationship with Stroke Risk Group

To understand if having difference in haemoglobin level values has any statistically significant impact on having difference in risk of stroke, we perform a hypothesis test.

**Null Hypothesis (H0):** Having different haemoglobin levels has no significant impact on having higher risk of stroke.

**Alternate Hypothesis (HA):** Having different haemoglobin levels has a significant impact on having higher risk of stroke.

We get the following result after performing One-Way ANOVA test:



Since, p-value > 0.05 (alpha), we fail to reject the Null Hypothesis (H0) and conclude that having difference in haemoglobin values has no statistically significant effect on having a difference in risk of stroke.

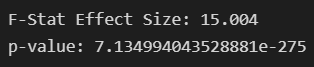
#### Diastolic Blood Pressure Relationship with Stroke Risk Group

To understand if having difference in Diastolic BP level values has any statistically significant impact on having difference in risk of stroke, we perform a hypothesis test.

**Null Hypothesis (H0):** Having different diastolic BP levels has no significant impact on having higher risk of stroke.

**Alternate Hypothesis (HA):** Having different diastolic BP levels has a significant impact on having higher risk of stroke.

We get the following result after performing One-Way ANOVA test:



Since, p-value < 0.05 (alpha), we reject the Null Hypothesis (H0) and conclude that having a difference in diastolic BP values has a statistically significant effect on having a difference in risk of stroke.

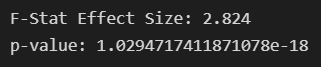
#### Pulse Rate Relationship with Stroke Risk Group

To understand if having difference in Pulse Rate values has any statistically significant impact on having difference in risk of stroke, we perform a hypothesis test.

**Null Hypothesis (H0):** Having different Pulse Rate values has no significant impact on having higher risk of stroke.

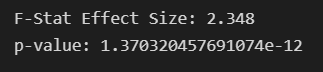
**Alternate Hypothesis (HA):** Having different Pulse Rate values has a significant impact on having higher risk of stroke.

We get the following result after performing One-Way ANOVA test:



Since, p-value < 0.05 (alpha), we reject the Null Hypothesis (H0) and conclude that having a difference in Pulse Rate values has a statistically significant effect on having a difference in risk of stroke.

We also checked the same hypothesis with the 2nd Pulse Reading that was provided in the dataset and came to the same conclusion. The results were:



# Data Exporting

We export our Dataset of individuals who are classified to be in higher risk of stroke category to an excel file named “output\_data\_stroke.xlsx” with the following features:

* Unique ID
* Sex
* District

We see that our dataset is classified successfully in accordance to our Project objective so we decide to conclude our project here.