STROKE PREDICTION AND ANALYSIS

Abhishek K

Problem Statement:

A stroke occurs when the blood supply to part of your brain is interrupted or reduced, preventing brain tissue from getting oxygen and nutrients. Brain cells begin to die in minutes. A stroke is a medical emergency, and prompt treatment is crucial. Early action can reduce brain damage and other complications. According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths.

This work aims at analyzing the patterns present in the dataset and predicting if a person is likely to encounter stroke, with the help of attributes such as hypertension, average glucose level, marital status, smoking status present in the dataset.

Scope:

The chosen dataset is procured and maintained by a private health organization, which pertains only to a certain population. There are quite a few inconsistencies in the dataset which have to be rectified in the data preprocessing step, in order to get accurate results.

The dataset is not balanced (high imbalance), where the total number of data points with the label as ‘stroke’ =0 is significantly higher than that of the data points with the label as ‘stroke’ =1, and due to the imbalance, the dataset is biased towards data points with target label as 0.

The dataset consists of relatively fewer data points, which degrades the performance of the classification models.

All analysis statistics and performance metrics of various models are restricted to a particular population represented by the dataset.

Dataset description:

1) id: The primary and 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) Residence_type: "Rural" or "Urban"

7) work_type: "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"

8) ever_married: "No" if not married or "Yes" if married

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 has a stroke or 0 if not

Categorical variables such as Residence_type, work_type, smoking_status are encoded using the dummy encoder (one hot encoder) method, to make it suitable for training and testing.

And variables such as ever_married, Residence_type are encoded since their values are restricted to distinct two values (eg: ever_married: "No" or "Yes")

Proposed solution:

Exploratory Data Analysis (EDA)

Under this section, various kinds of graph visualizations of various parameters have been implemented to analyze the unique patterns and relationships among different variables present in the dataset.

Data cleaning and preprocessing

The raw dataset couldn’t be used as such, as there were a lot of inconsistencies and empty rows. Hence they have been processed and transformed into an entity suitable for classification.

The categorical variables such as work_type have to be dummy encoded in order to consider that attribute as a feature, as Machine learning algorithms don't take strings, categories as such.

Hypothesis testing

Chi-square test has been employed to check the dependency of variables on other variables, typically a feature variable and a target variable are considered. Hypertension and stroke variables are tested for dependency.

ML classification models used :

- Logistic Regression
- Decision Tree Classifier
- XGBoost Classifier
- Extra Tree Classifier

Performance metrics used :

- Precision
- Recall
- F1 score
- Accuracy

Data imbalance :

Since the dataset has more number of data points corresponding to target = 0 than that of the data points corresponding to target = 1, the dataset is biased towards the target value = 0.

Hence, the dataset is over-sampled using SMOTE technique in which the number of data points belonging to the minority class is increased.

ML classifier models are implemented and performance metrics are evaluated for the oversampled dataset.

Result :

Performance metrics of models run on the unsampled dataset is compared to that of models run on the sampled dataset and the best classifier is chosen.

Tools used

- Python (Anaconda navigator)
- Jupyter environment (VS code)
- imblearn
- Seaborn
- Numpy
- Pandas
- Matplotlib
- Scipy

Other similar tools

- PyCharm
- Keras
- Tensorflow
- Pytorch

Performance metrics:

Evaluating the performance of various classifiers helps us to make informed decisions as to which classifier to be chosen as the best.

Performance metrics considered :

- Accuracy: It is the ratio of the total number of correct predictions to the total number of predictions made. Accuracy provides us with an overall performance rather than performance specific to a target class.

- Precision: It is the ratio of the total number of true positives to the total number of data points with ground values as positive.

- Recall: It is the ratio of the total number of true positives to the total number of data points with predicted values as positive.

- F1- score: Harmonic mean of recall and precision.

Exploratory Data Analysis

Data cleaning and preprocessing

Oftentimes, many feature columns will be left with empty values, ambiguous values which have to be either removed or filled with the average/median value. In this dataset, the bmi column had around 200 null values, which was later filled with the average value of the rest of the entries.

Feature columns such as gender, work_type, smoking_status are categorical variables that have more than two unique values and were transformed into binary-valued features to make the dataset suitable for using it to train and test ML models.

One hot encoder technique was used to perform the above-mentioned feature transformation.

Feature columns such as gender, ever_married, Residence_type have only two unique values which can be easily mapped to binary 0 and 1 values. Label encoder technique was used to transform binary string-valued columns into binary 0 and 1 valued columns.

Pearson correlation metric

Pearson correlation metric has been used to find the degree of correlation between various feature attributes and target labels.

Pearson correlation is calculated as : covariance(x,y) / (std_dev(x) * std_dev(y))

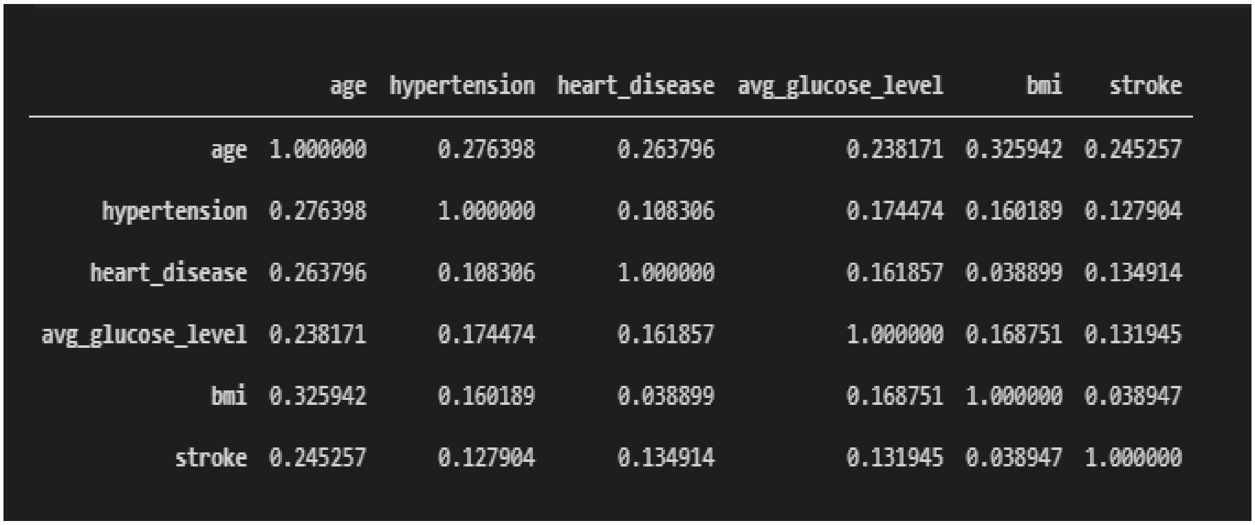
Where,

x is any feature attribute and y is the target attribute.
std_dev represents the standard deviation function.

Pearson correlation for each feature with respect to the target attribute is given as :

```
'age': 0.2452573461709748,
'hypertension': -0.12790382346648038,
'heart_disease': -0.13491399696869283,
'ever_married': 0.20833974165701019,
'Residence_type': -0.015457965477256957,
'avg_glucose_level': 0.13194544082571016,
'bmi': 0.03894659651202003,
'encoded_Female': -0.009026602170652232,
'encoded_Male': 0.009117154023008658,
'encoded_Other': -0.003166422855664966,
'encoded_Govt_job':  0.0026767045331463485,
'encoded_Never_worked':  -0.014882457537542752,
'encoded_Private':  0.011888235156031489,
'encoded_Self-employed':  0.062168257271196445,
'encoded_children':  -0.08386926606516618,
'encoded_Unknown':  -0.055891709388241236,
```

'encoded_formerly smoked': 0.06455557529819726,
'encoded_never smoked': -0.004128687879042863,
'encoded_smokes': 0.008939203206269973



	age	hypertension	heart_disease	avg_glucose_level	bmi	stroke
age	1.000000	0.276398	0.263796	0.238171	0.325942	0.245257
hypertension	0.276398	1.000000	0.108306	0.174474	0.160189	0.127904
heart_disease	0.263796	0.108306	1.000000	0.161857	0.038899	0.134914
avg_glucose_level	0.238171	0.174474	0.161857	1.000000	0.168751	0.131945
bmi	0.325942	0.160189	0.038899	0.168751	1.000000	0.038947
stroke	0.245257	0.127904	0.134914	0.131945	0.038947	1.000000

Negative value: attributes are inversely proportional.
Positive value: attributes are directly proportional.
Zero: attributes are not dependent on each other.

Hypothesis Testing:

A chi-square test is done to check for independence between two variables, namely hypertension, and stroke where hypertension is a feature variable and stroke is a target variable.

Null hypothesis: hypertension and stroke are dependent variables.
Alternative hypothesis: hypertension and stroke are not dependent variables.

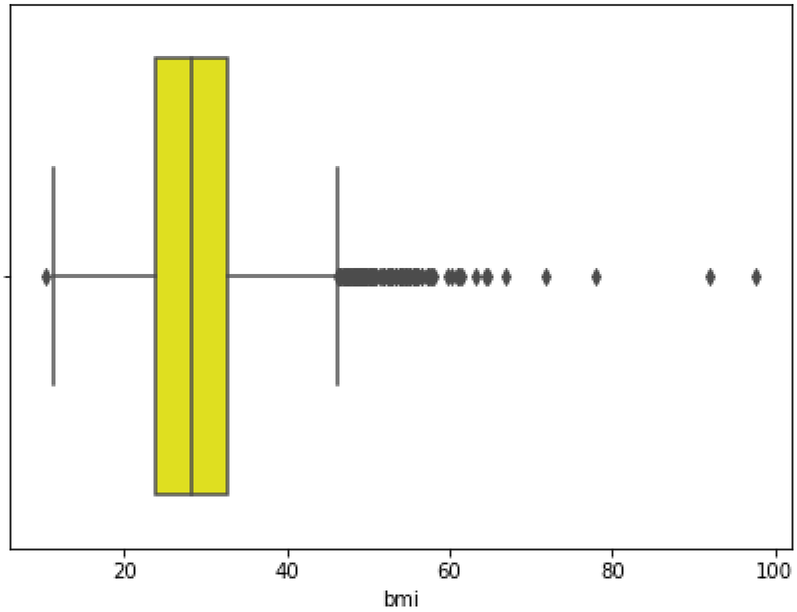
Level of significance: 0.01 (1%)

variable	stroke = 1	stroke = 0
hypertension = 1	66	432
hypertension = 0	183	4429

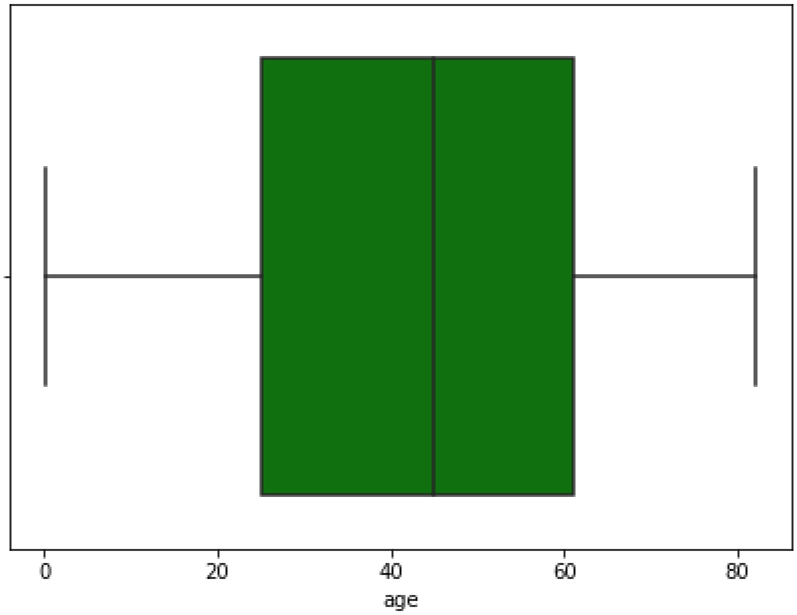
p = 1.661621901511823e-19

Since the p-value is less than 0.01, the null hypothesis is rejected. Hence, hypertension and stroke are not dependent variables.

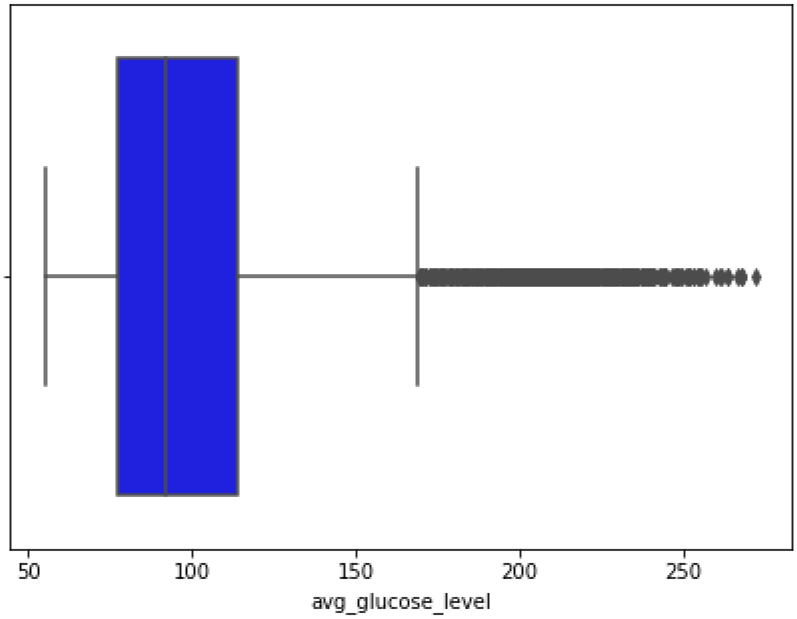
BOXPLOTS:



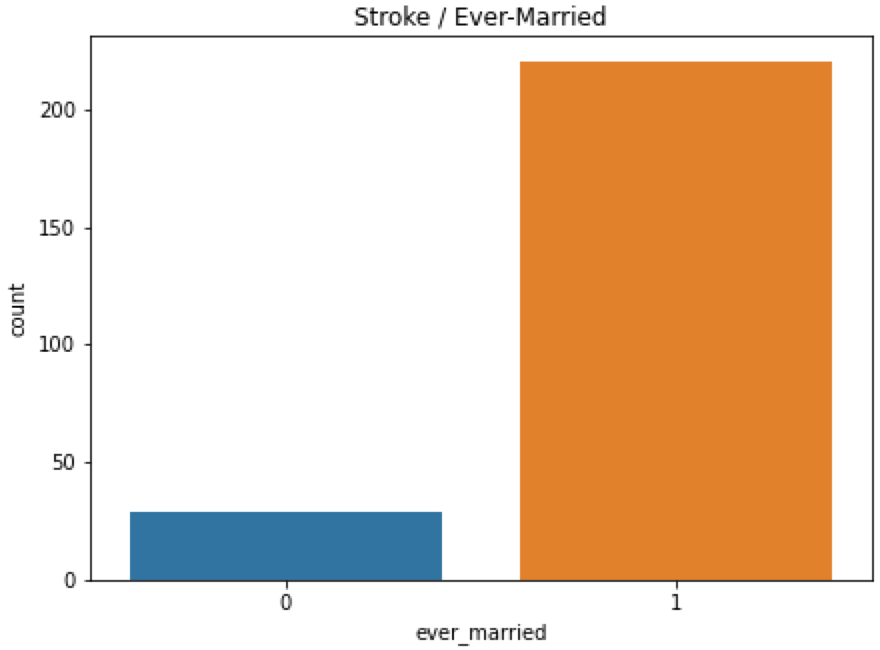
This plot shows that the median bmi of people with stroke is around 28



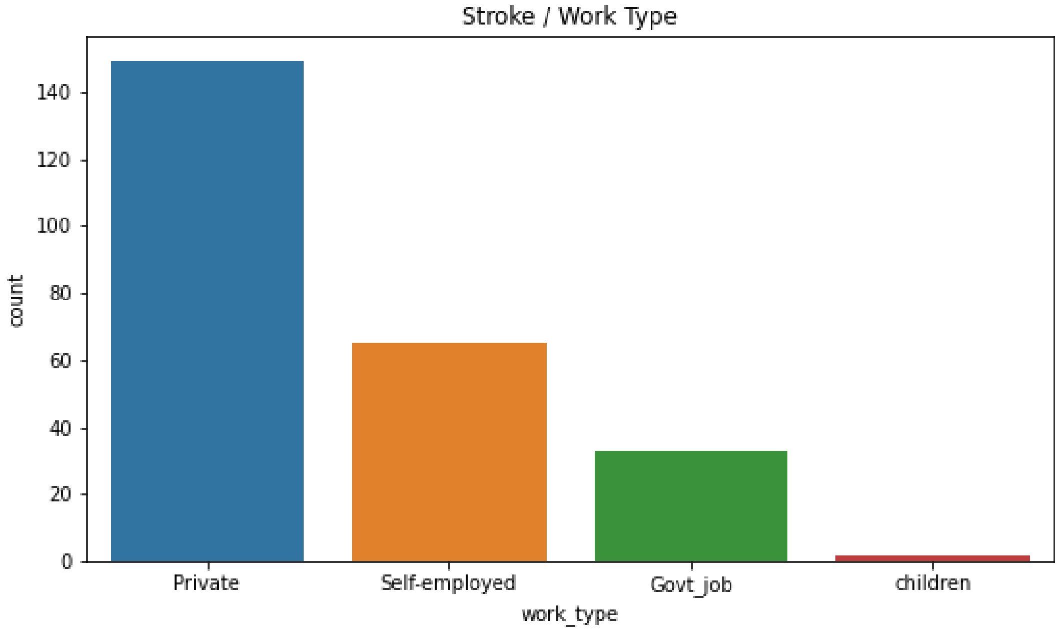
This plot shows that median age at which people encounter stroke is around 45 -50



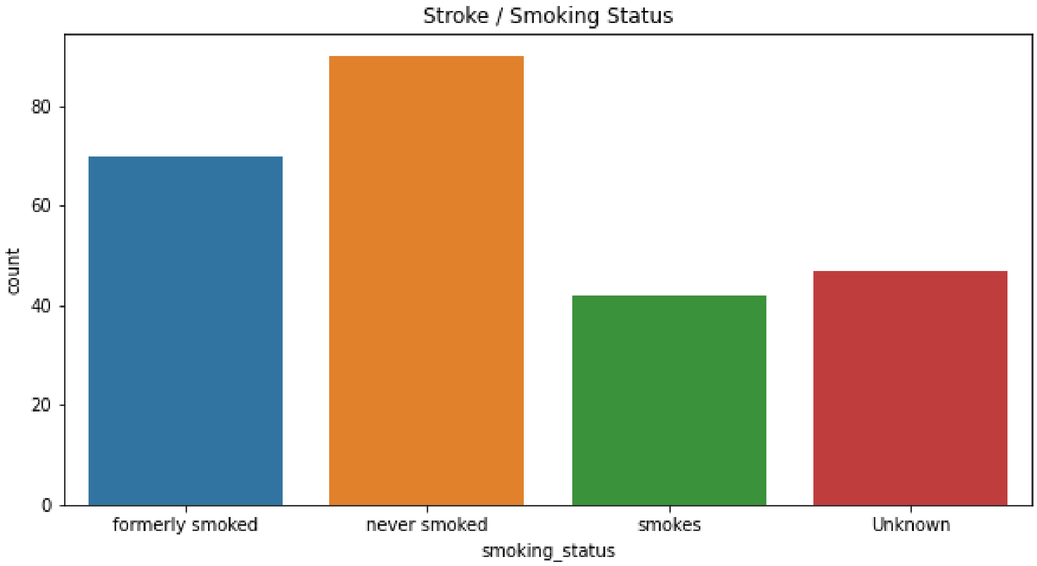
This plot shows that the median glucose level in people with stroke is around 95-100



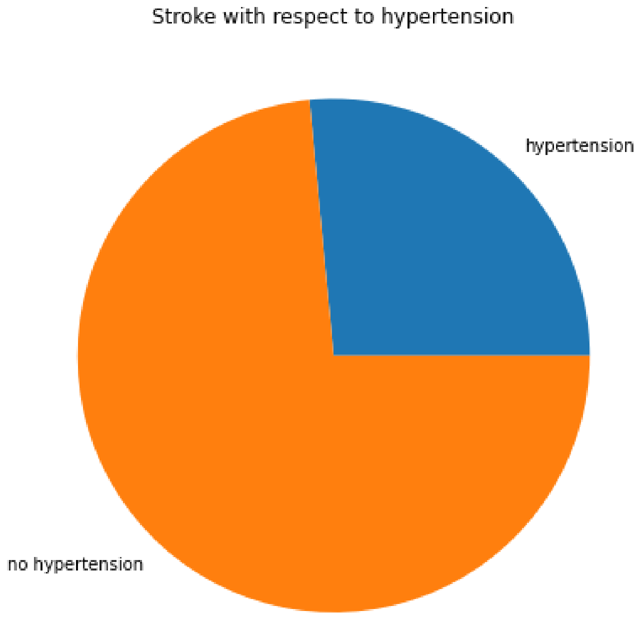
This plot shows that a person’s marital status and the probability of encountering stroke are highly correlated. Married individuals showed a high chance of encountering stroke.



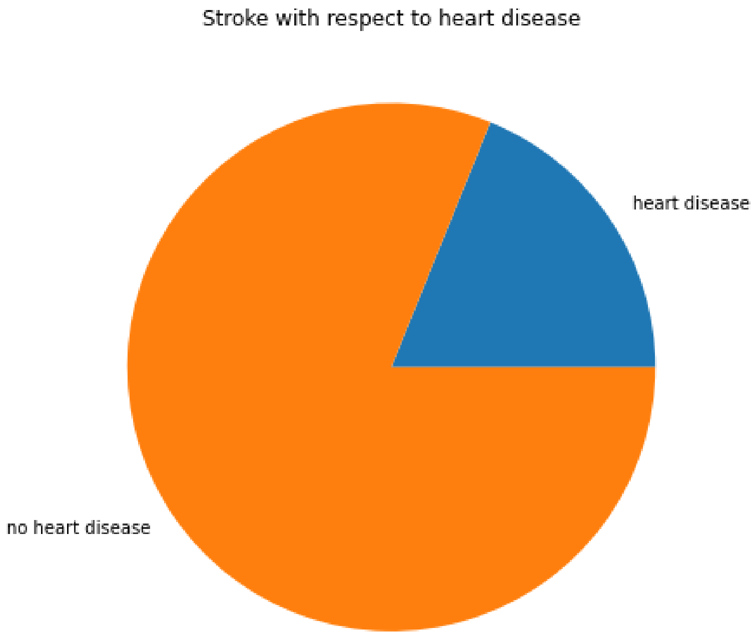
This plot shows that people working in the private sector are highly susceptible to encountering strokes.



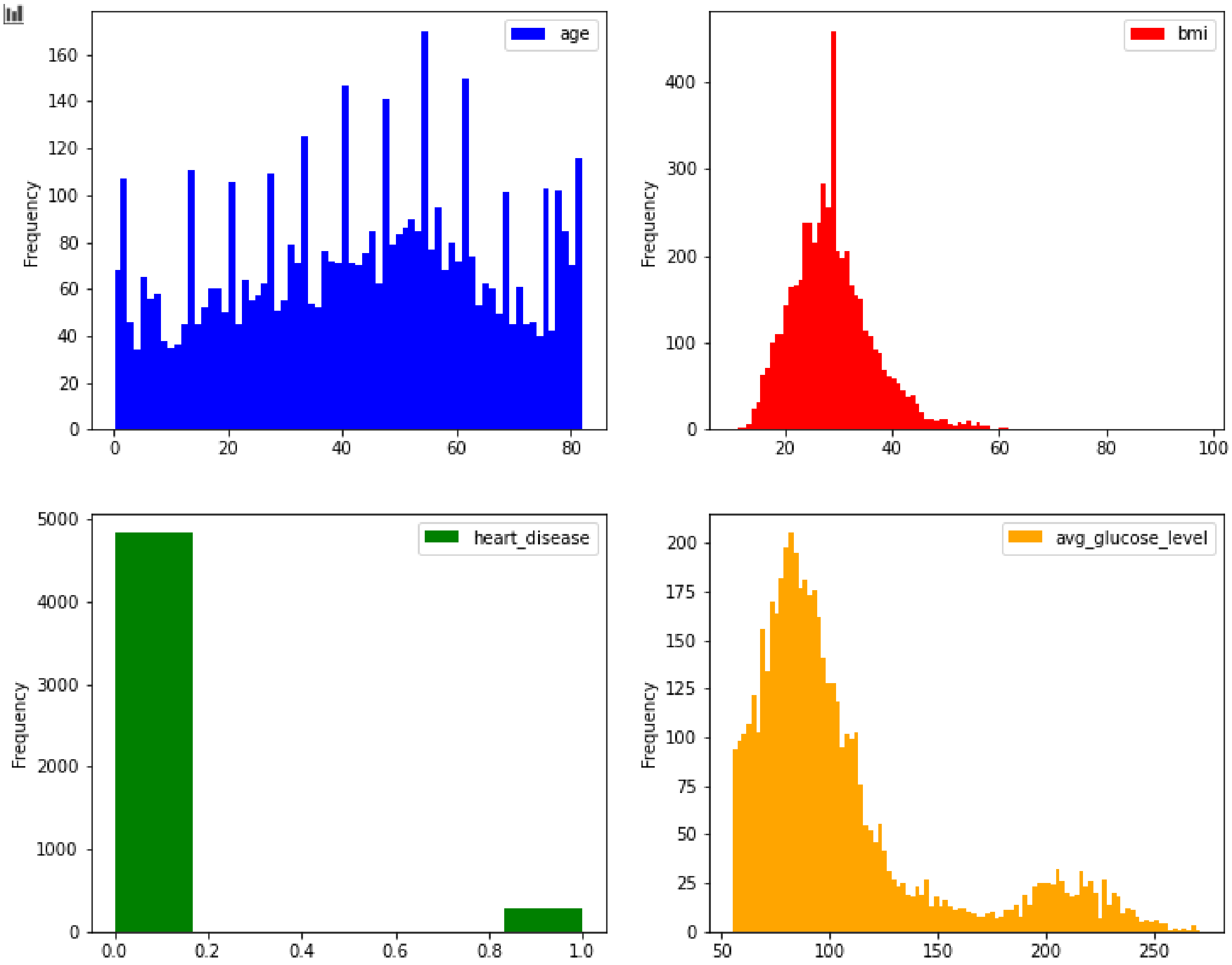
This plot shows that people who have smoked in the past and smoke currently are prone to stroke.



This pie chart shows that people with no hypertension have a higher risk of stroke than people with hypertension.

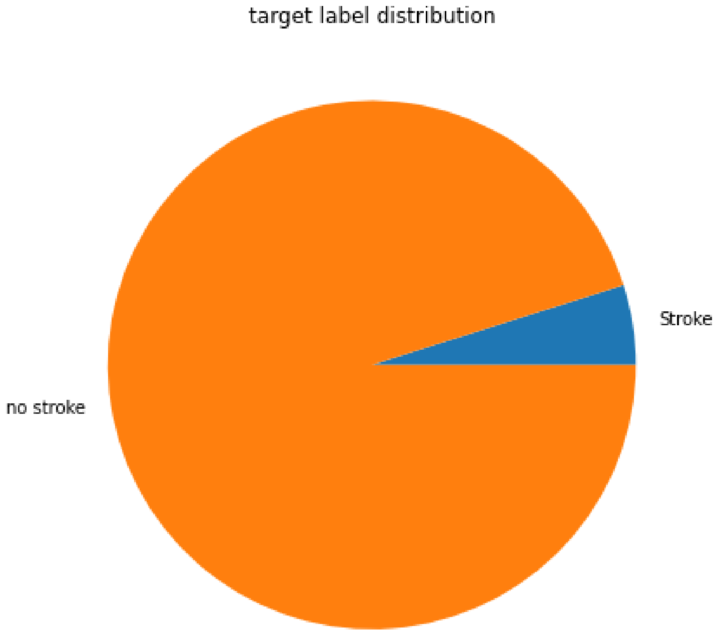


This pie chart shows that people with no heart disease have a higher risk of encountering stroke, than people with heart disease.



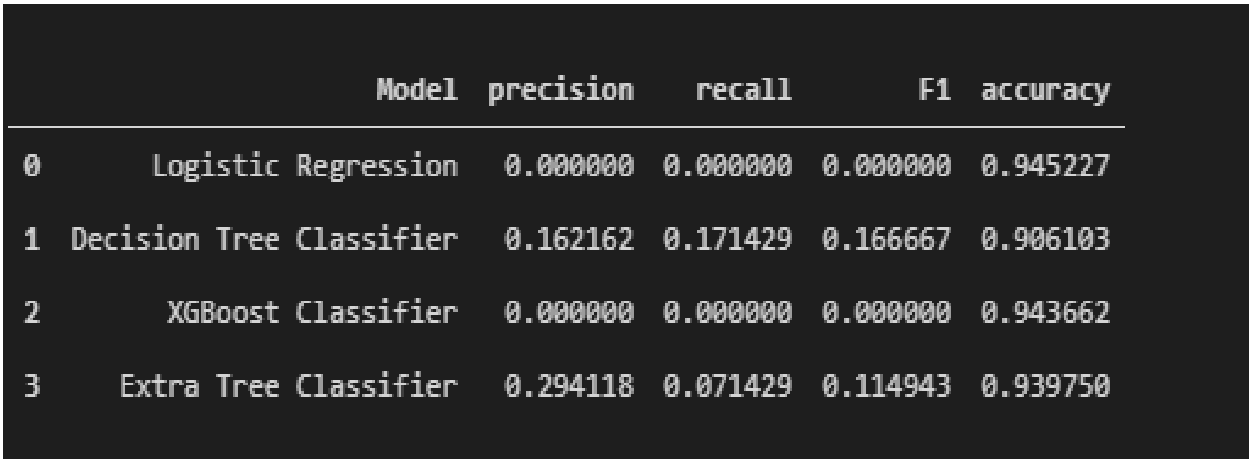
This frequency plot shows that:

- People belonging to the age group of 45- 55 have encountered strokes most frequently.
- People with BMI values ranging from 28 -30 have encountered strokes most frequently.
- People with glucose levels ranging from 75 - 80 have encountered strokes most frequently.



This pie chart depicts the huge imbalance in the dataset, where the number of data points with the target as 1 is significantly lesser than the number of data points with the target as 0. Hence this imbalance makes the dataset skewed and biased towards one particular target class.

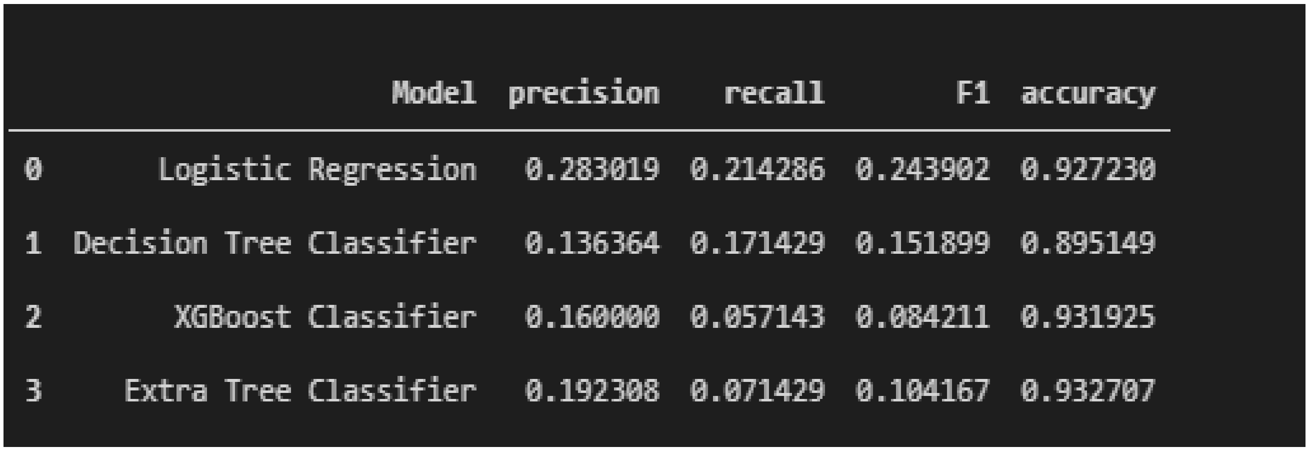
Result:



```
                     Model  precision    recall        F1  accuracy
0      Logistic Regression   0.000000  0.000000  0.000000  0.945227
1  Decision Tree Classifier   0.162162  0.171429  0.166667  0.906103
2        XGBoost Classifier   0.000000  0.000000  0.000000  0.943662
3     Extra Tree Classifier   0.294118  0.071429  0.114943  0.939750
```

The above mentioned metric scores correspond to the unsampled data, where the Number of 1 = 249 and number of 0 = 4861. Hence all the classifier models would have been trained better to classify class 0 and neglect class 1.

Therefore, this evaluation score cannot be considered for selecting the best classifier model. We need to implement sampling techniques to balance the dataset.



```
                     Model  precision    recall        F1  accuracy
0      Logistic Regression   0.283019  0.214286  0.243902  0.927230
1  Decision Tree Classifier   0.136364  0.171429  0.151899  0.895149
2        XGBoost Classifier   0.160000  0.057143  0.084211  0.931925
3     Extra Tree Classifier   0.192308  0.071429  0.104167  0.932707
```

Now that we have sampled the dataset, all the classifiers have shown better performance through their metric scores. It is evident from the above table that the Logistic Regression model has performed better than the rest of the models in terms of precision, recall, and F1 score and with a decent accuracy score as such.

Conclusion:

The main objective of this work is to predict if a person is likely to encounter stroke, by considering various factors and health parameters associated with the person. The whole process begins with data analysis, searching for relationships, patterns across various attributes, and then the data is cleaned and preprocessed to make the data suitable for prediction.

Since the data is imbalanced, we implement sampling techniques to counteract the imbalance, and then classifiers are trained on the sampled dataset, which yields better results than the models that run on unsampled data.

After considering various performance metrics and analysis statistics, Logistic regression has performed better than other models. Hence logistic regression can be used to classify if a person is likely to encounter stroke or not, which is purely specific to any particular dataset and population.

This work can be further scaled up and produce better performance scores with more amount of clean data and resources.

References

- M. S. Singh and P. Choudhary, "Stroke prediction using artificial intelligence," 2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON), Bangkok, Thailand, 2017, pp. 158-161, DOI: 10.1109/IEMECON.2017.8079581.

- Amini L, Azarpazhouh R, Farzadfar MT, Mousavi SA, Jazaieri F, Khorvash F, Norouzi R, Toghianfar N. Prediction and control of stroke by data mining. Int J Prev Med. 2013 May;4(Suppl 2): S245-9. PMID: 23776732; PMCID: PMC3678226.

- Przelaskowski A, Sklinda K, Bargieł P, Walecki J, Biesiadko-Matuszewska M, Kazubek M. Improved early stroke detection: Wavelet-based perception enhancement of computerized tomography exams. *Comput Biol Med.* 2007;37:524–33. [PubMed] [Google Scholar]