
BIA
BOSTON
INSTITUTE OF
ANALYTICS
®

PROBLEM STATEMENT

We have a health care data about stroke with some featured columns. We have to predict which features are highly related to the chance of having a stroke.

Data Science Life Cycle

1. Data Understanding
2. Tools used
3. Data Pre-processing
4. EDA
5. Building Model

Data Understanding

- Data consists of 12 features with 5110 records
- 3.93% data is missing from BMI feature
- There are no duplicate entries

The dependent feature in our data set in ‘Stroke’. Since, it is discrete type of data and the output is binary, we can get the best result if we apply classification.

Tools Used

- Jupyter Notebook
- Python
- Pandas
- Matplotlib
- Seaborn
- Scikit-learn
- IMBlearn
- Power BI

About the Dataset

1) id: 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) ever_married: "No" or "Yes"
7) work_type: "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
8) Residence_type: "Rural" or "Urban"
9) avg_glucose_level: Average Glucose Level in Blood which can indicate diabetes
10) bmi: Body Mass Index
11) smoking_status: "formerly smoked", "never smoked", "smokes" or "Unknown"*
12) stroke: patient had a stroke (1 : Yes, 0 : No)

Data Preprocessing

1. Handling Missing Values –
   The dataset had 201 missing records of BMI which is equivalent to 3.93% of the total Data. Hence, we can fill it with mean or median. Since the data distribution is Skewed, we will used median.

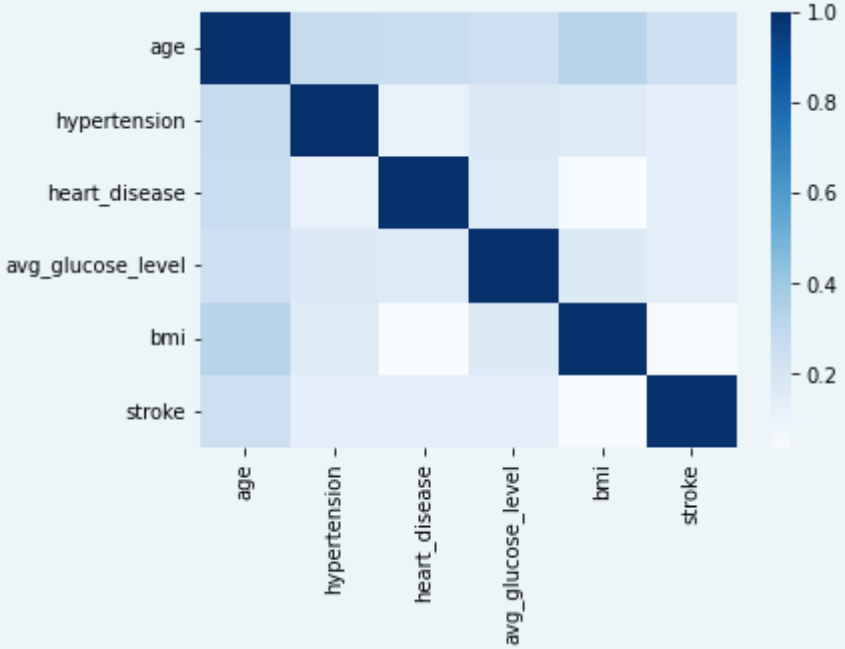
2. Handling Improper Data Types –
   The age feature is given as float type, we can convert it to int64 for better performance
   and to build memory efficient model.

Feature Selection

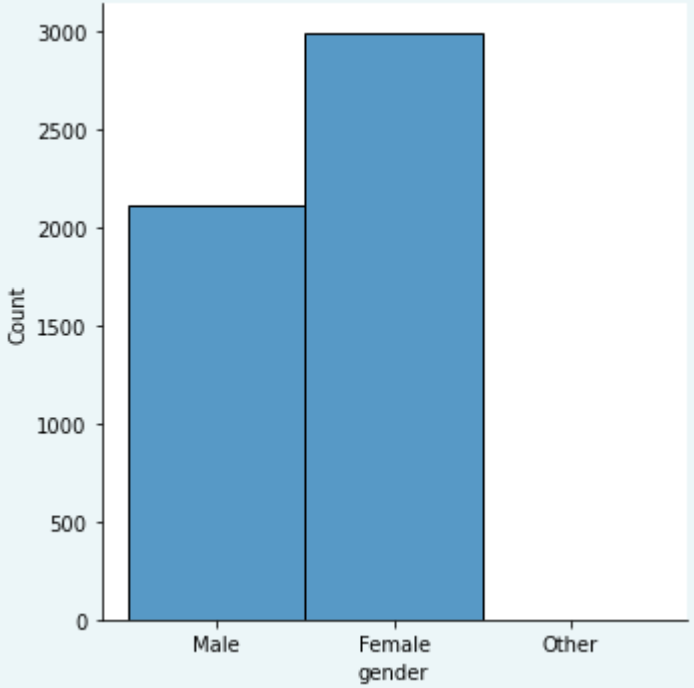
We have an id column in our dataset. This feature is to identify unique record. We don't require such features
as it does not provide us any information related to the factor affecting the person having stroke or not.

EDA

Correlation between numerical features

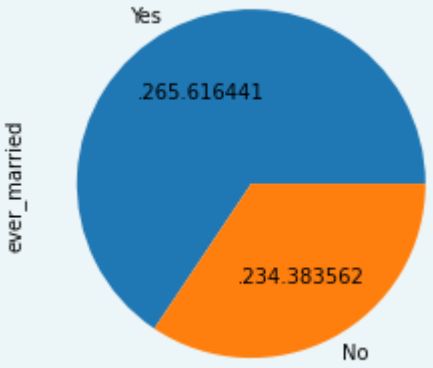


Distribution of Gender in the data

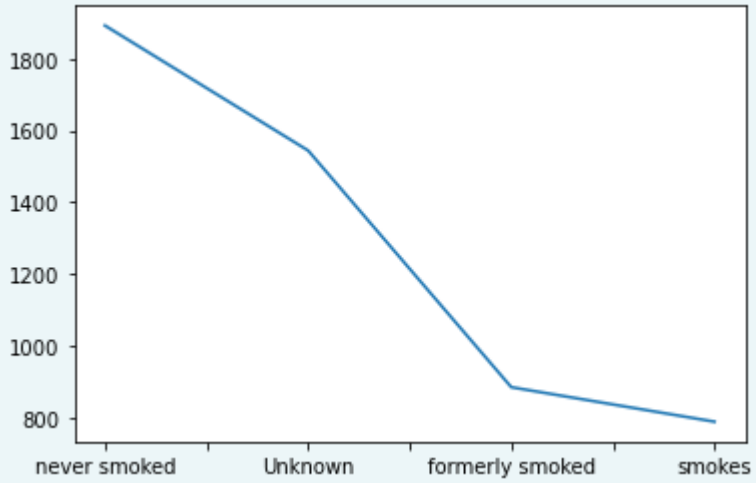


EDA

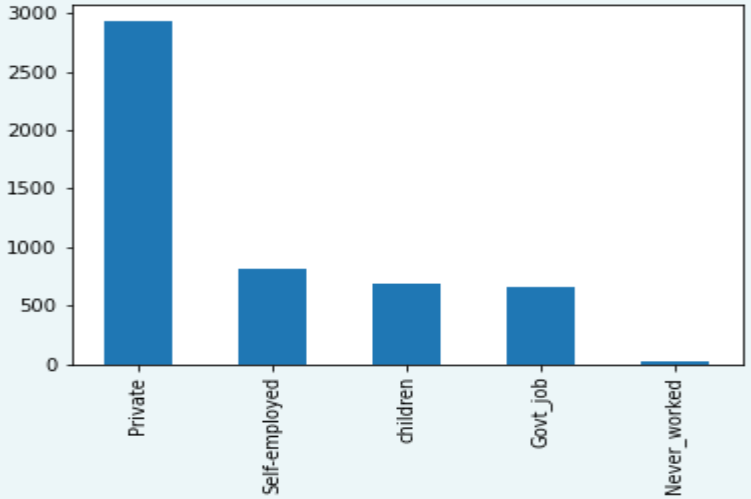
Distribution of Martial Status



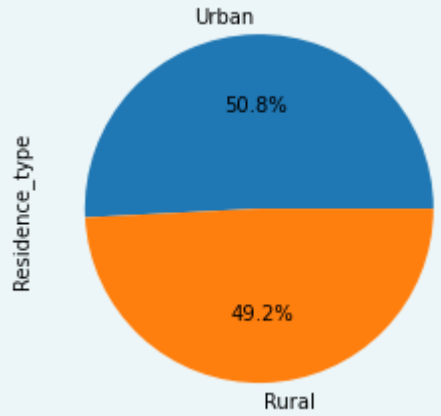
Distribution of Smoking Habit



Distribution of Work Type

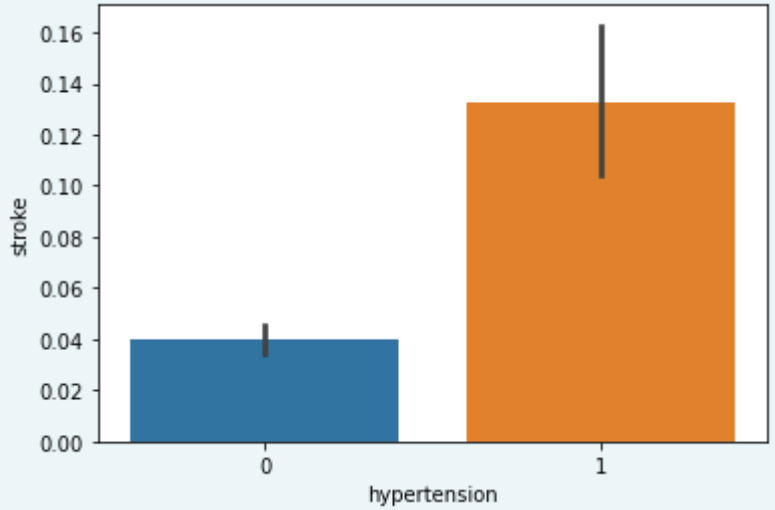


Distribution of Residence Type

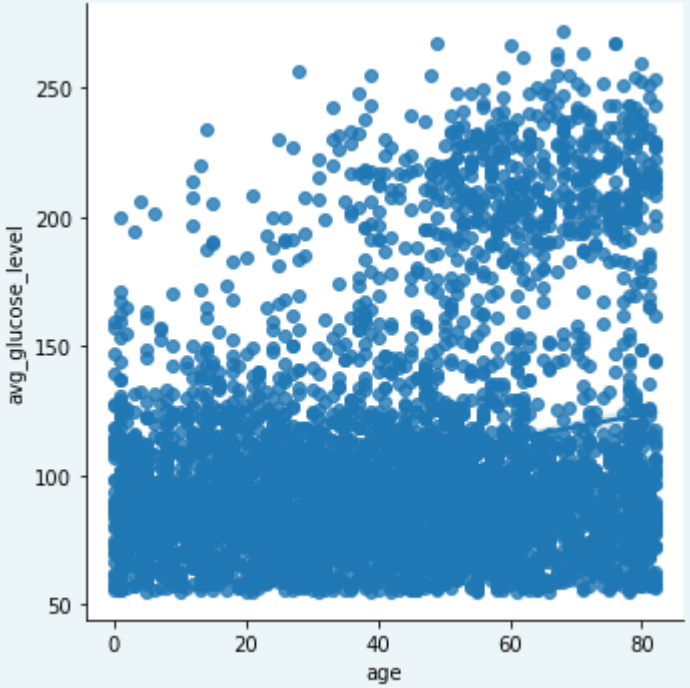


EDA

Count of Hypertension with Stroke

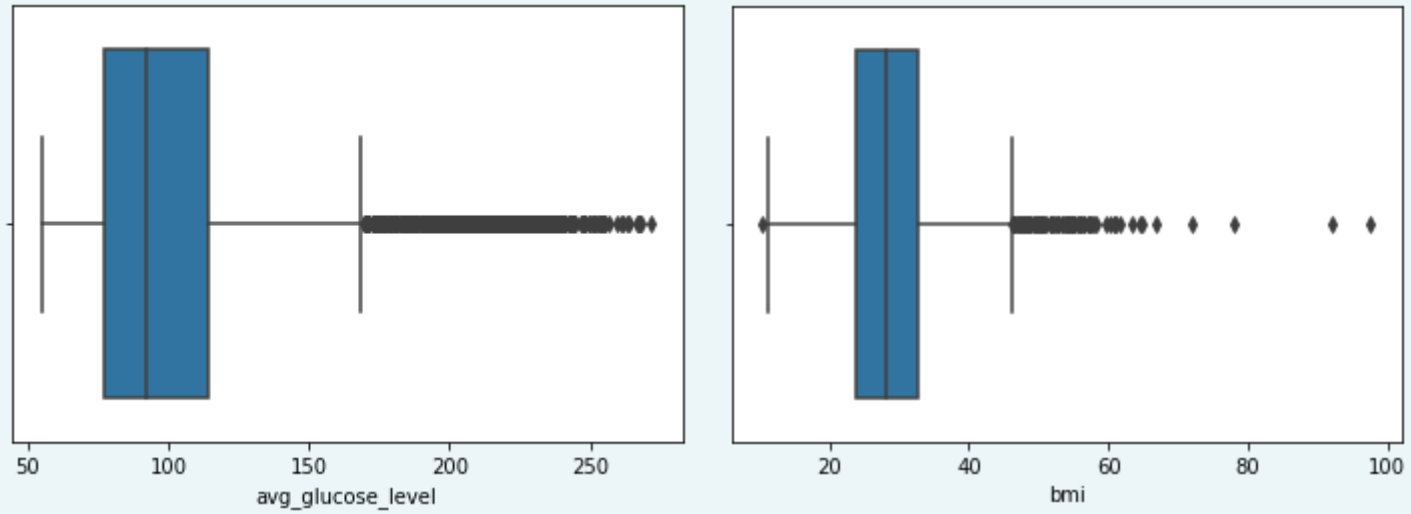


Scatter of Glucose level with Age



EDA

Outliers



Average Glucose Level

Body Mass Index

Since, these outliers are the data points which will help us to detect the stroke, we will not process anything on these outliers.

Building the Model

Encoding

Dataset consists of object type values. We need to convert those values in numeric type. Hence, we use Label Encoder.

Splitting the Data

Our aim is to detect the stroke. In order to train our model, we need to differentiate the columns as X and Y. Further we divided the data into 75-25 % to train and test. 75% of data will go for training and 25% will be used for evaluation.

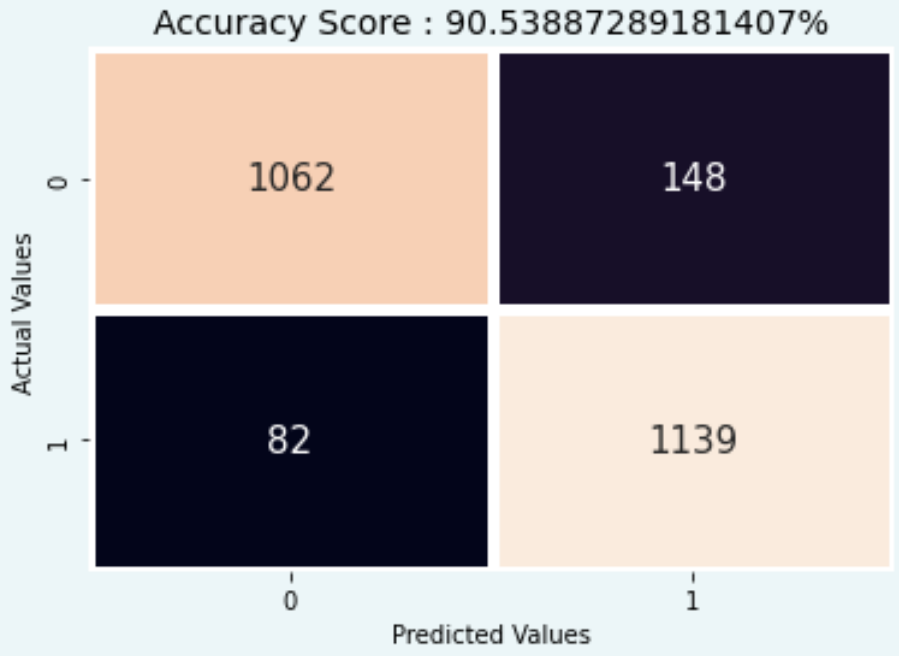
SMOTE Oversampling Technique

The balance between the output Yes and No had a huge difference. From 5110 records only 249 people had stroke. This can create a huge bias in our model. Hence, we use oversampling to avoid this imbalance.

Building the Model

Building with Decision Tree

- Decision trees are a simple yet powerful model used in machine learning.
- They work by recursively splitting the data into smaller subsets based on the features that provide the most information gain.
- Each internal node of a decision tree represents a decision based on a feature, and each leaf node represents a class label or numerical value.

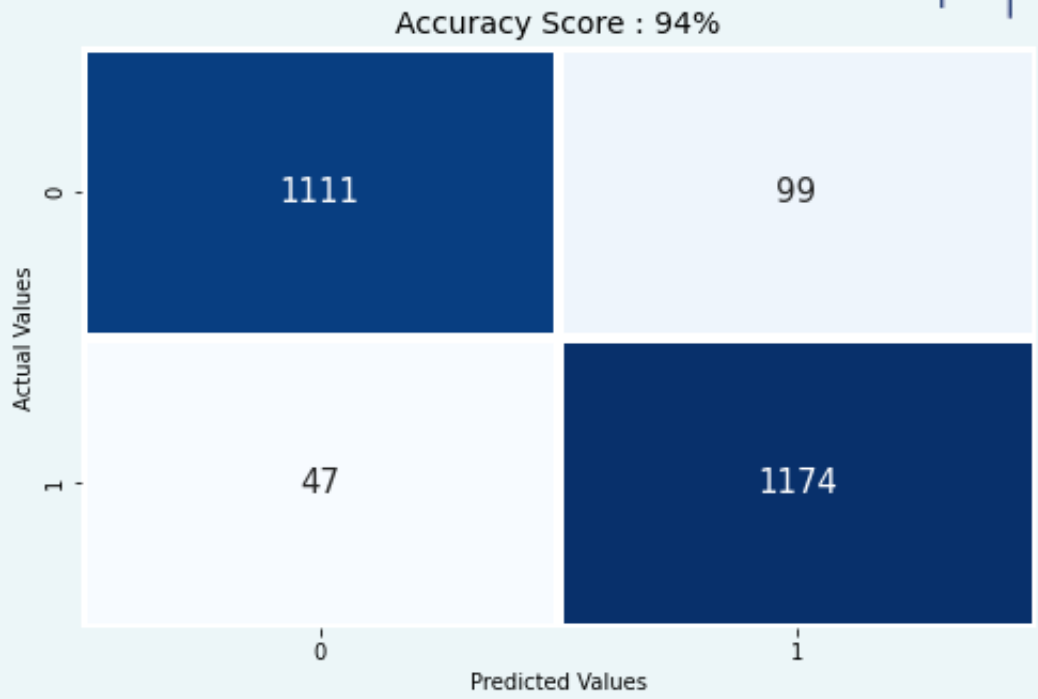


Since we achieved accuracy score of 90.5% , we will try to build using different algorithm.

Building the Model

Building with Random Forest

•Random Forest is an ensemble of decision trees, where each tree is trained on a randomly sampled subset of the data and a random subset of the features.
•This randomness helps to prevent overfitting and increases the model's robustness and accuracy.
•The final prediction is made by aggregating the predictions of all the individual trees.



Using this technique, we have achieved accuracy score of 94% , Let us use this model to check the feature importances.

Evaluating the Model

	Features	Importance
1	age	0.395322
2	avg_glucose_level	0.203436
3	bmi	0.164148
4	work_type	0.074471
5	smoking_status	0.055420
6	Residence_type	0.029103
7	gender	0.028290
8	ever_married	0.022530
9	hypertension	0.015983
10	heart_disease	0.011298

Dashboard

Input Filters – Age, Residence Type, Gender

Conclusion

Using this model and Feature Importances, we can say that with increasing age, the chances of strokes increases. Patients with diabetes also have chances of stroke. Obesity also has risk with Stroke.

The use of machine learning algorithms can assist healthcare professionals in making accurate and timely stroke predictions, which can lead to better patient outcomes and reduced healthcare costs. However, it is essential to consider the ethical and privacy implications of using sensitive patient data in machine learning algorithms. Overall, stroke prediction using machine learning algorithms is a promising area of research that has the potential to improve stroke prevention and patient care.

Thank You!

BIA BOSTON INSTITUTE OF ANALYTICS ®