MINI PROJECT

ON

HEALTHCARE: STROKE PREDICTION

TEAM MEMBERS : ASWIN, GRAHAM, JUSTIN

DATASET: *https://drive.google.com/open?id=1D-cCS2pmZ6IEn_wl6_F5x0Ic9ku7tGAE*

PROBLEM STATEMENT:

Stroke is a major global public health problem and also it is the leading cause of death and serious long-term disability. Our main goal is to determine the importance of the features on determining the stroke and predict whether a patient will have stroke or not.

DATASET DESCRIPTION

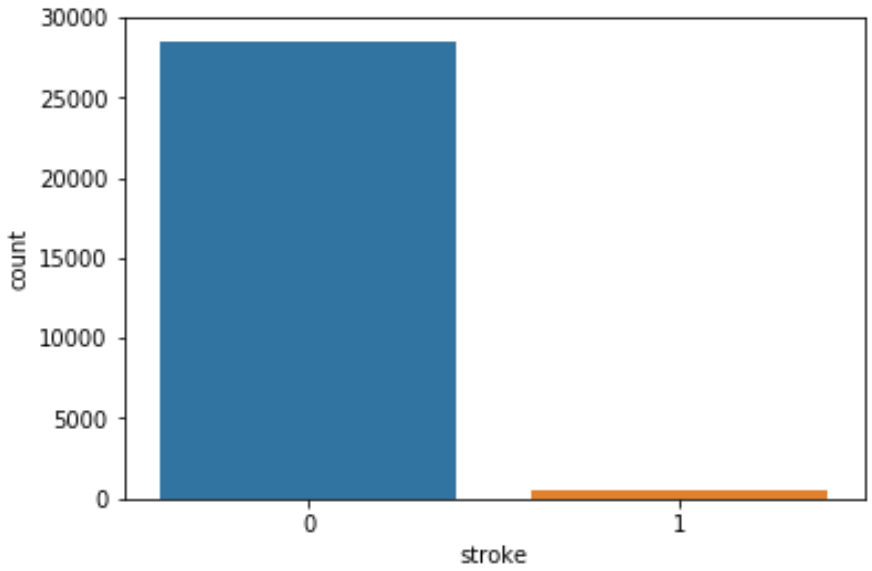
The dataset is obtained from kaggle.com.

The dataset contains (43400, 12) entries.

The following are the different attributes of the dataset

1	'id'	Patient id
2	'gender’	Gender of patient
3	'age'	Age of patient
4	'hypertension'	YES/NO
5	'heart_disease'	YES/NO
6	'ever_married'	YES/NO
7	'work_type'	'children', 'Private', 'Never_worked', 'Self-employed', 'Govt_job'
8	'Residence_type'	['Rural', 'Urban']
9	'avg_glucose_level'	Glucose level
10	'bmi'	Body mass Index
11	'smoking_status’	'never smoked', 'formerly smoked', 'smokes'
12	'stroke'	[0, 1]

The dataset is really challenging that it has got some issues which may affect the prediction. This dataset is highly imbalanced one as value counts in the stroke column has vast difference between them. so that the models may predict zero always.The preprocessing steps must be given more care before handling such unbalanced data. The following figure shows the value counts of our target variable “stroke”.



The dataset contains the following columns. Among these we try to find the relationship with “stroke”, which is the target variable.

```
data.columns
```

```
Index(['id', 'gender', 'age', 'hypertension', 'heart_disease', 'ever_married',
       'work_type', 'Residence_type', 'avg_glucose_level', 'bmi',
       'smoking_status', 'stroke'],
      dtype='object')
```

Another issue in the dataset is the number of missing values in the data set. The body mass index column (bmi) and smoke status columns contains missing values. Proper methods must be applied to resolve these issues .The following figure shows the counts of missing values corresponding to each variables.

```
data.isna().sum()
```

```
id                    0
gender                0
age                   0
hypertension          0
heart_disease         0
ever_married          0
work_type             0
Residence_type        0
avg_glucose_level     0
bmi                1462
smoking_status    13292
stroke                0
dtype: int64
```

It is important to check the datatypes of different variables in the data. The dataset contains categorical columns. The following figure shows the variables and their corresponding datatypes.

```
data.dtypes
```

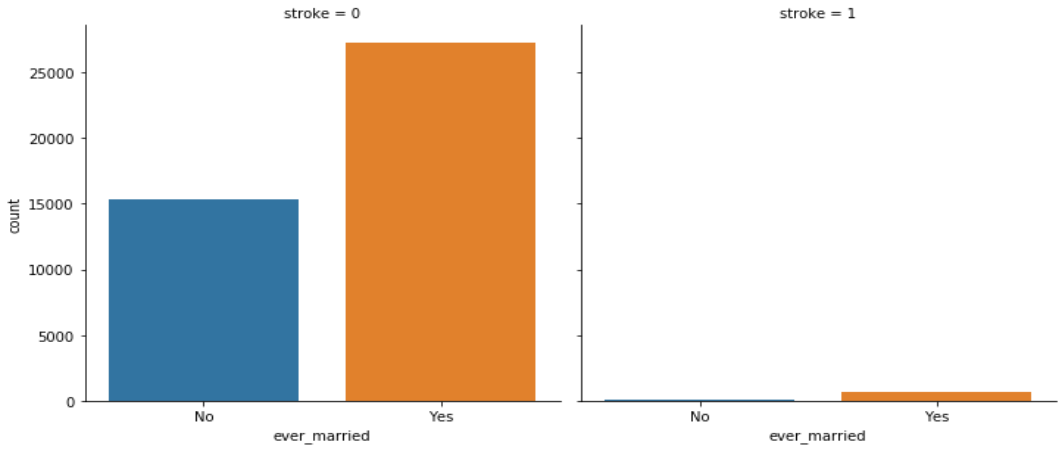
```
id                   int64
gender              object
age                float64
hypertension         int64
heart_disease        int64
ever_married        object
work_type           object
Residence_type      object
avg_glucose_level  float64
bmi                float64
smoking_status      object
stroke               int64
dtype: object
```

Good exploratory data analysis must be carried out to find more relationships and irregularities in the data.

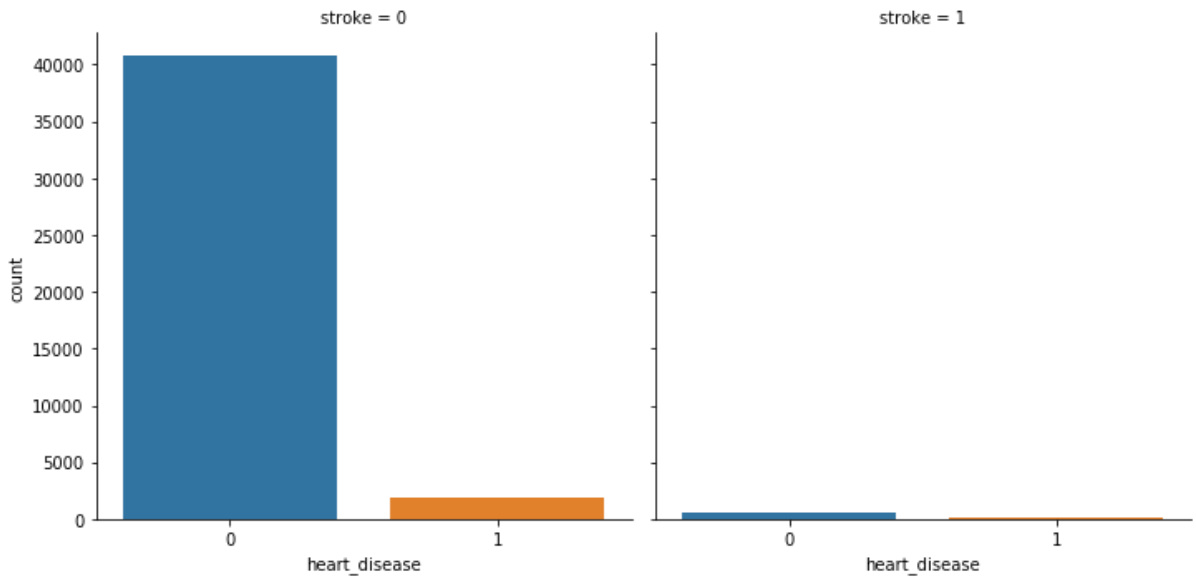
Exploratory data analysis

The dataset contains 43400 entries and 12 columns including the target variable “stroke”. Among these almost 1.8% have stroke. Our aim is to identify the most important features which determine whether a person has stroke. We here, plot every variable against stroke to find the relationships. **The prior assumptions are age , bmi, smoking status, hypertension and work status highly contribute to stroke.**

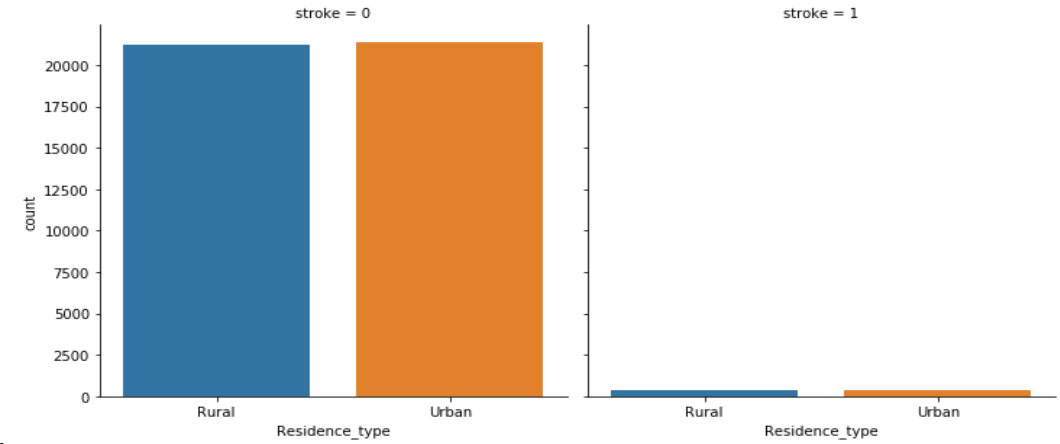
Let’s check the visualisations for getting better understanding.



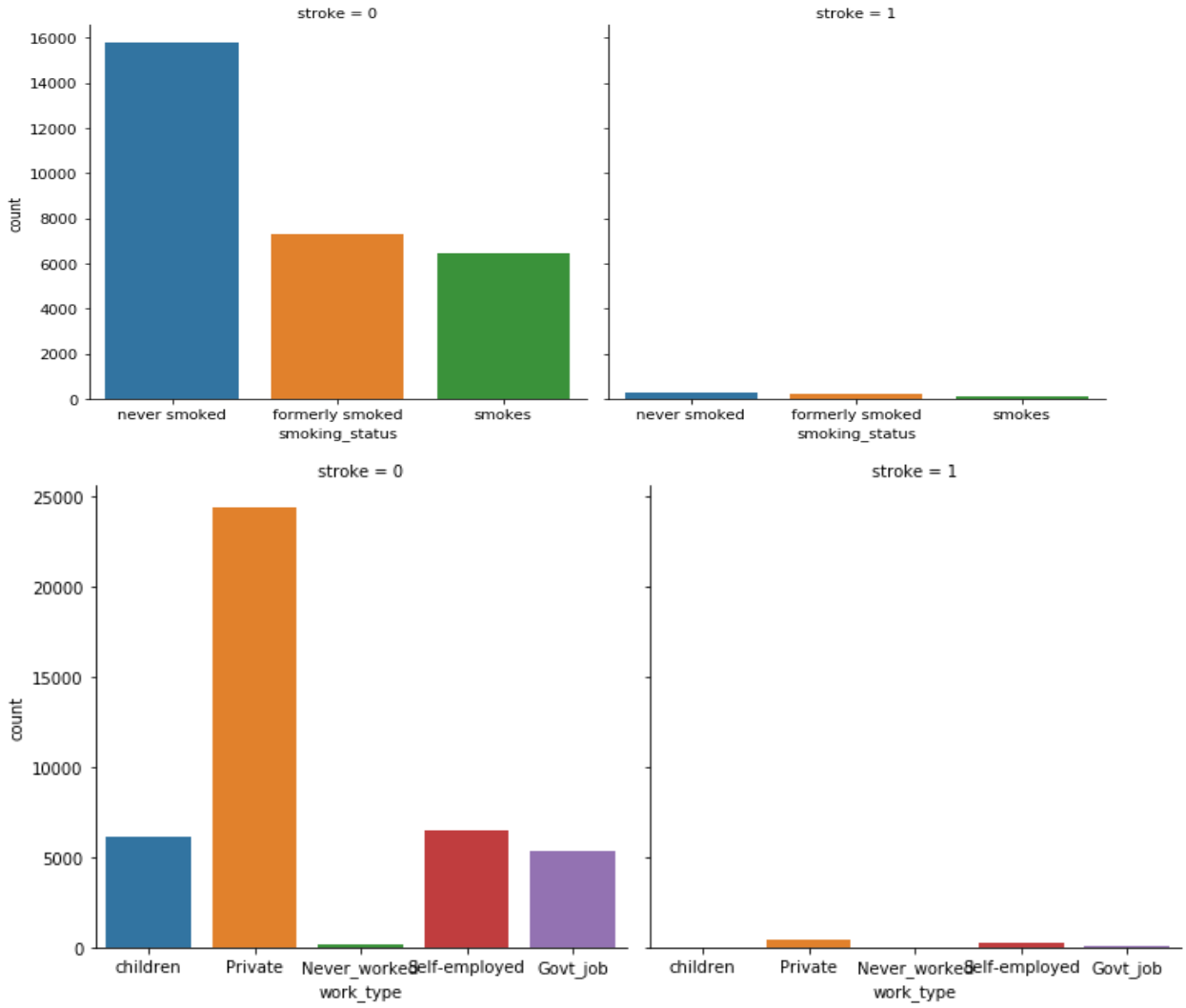
From the graph we can infer that the relationship between the variables ever_married and stroke is very week.



From the visualization no inferences can be made. Further investigation is needed

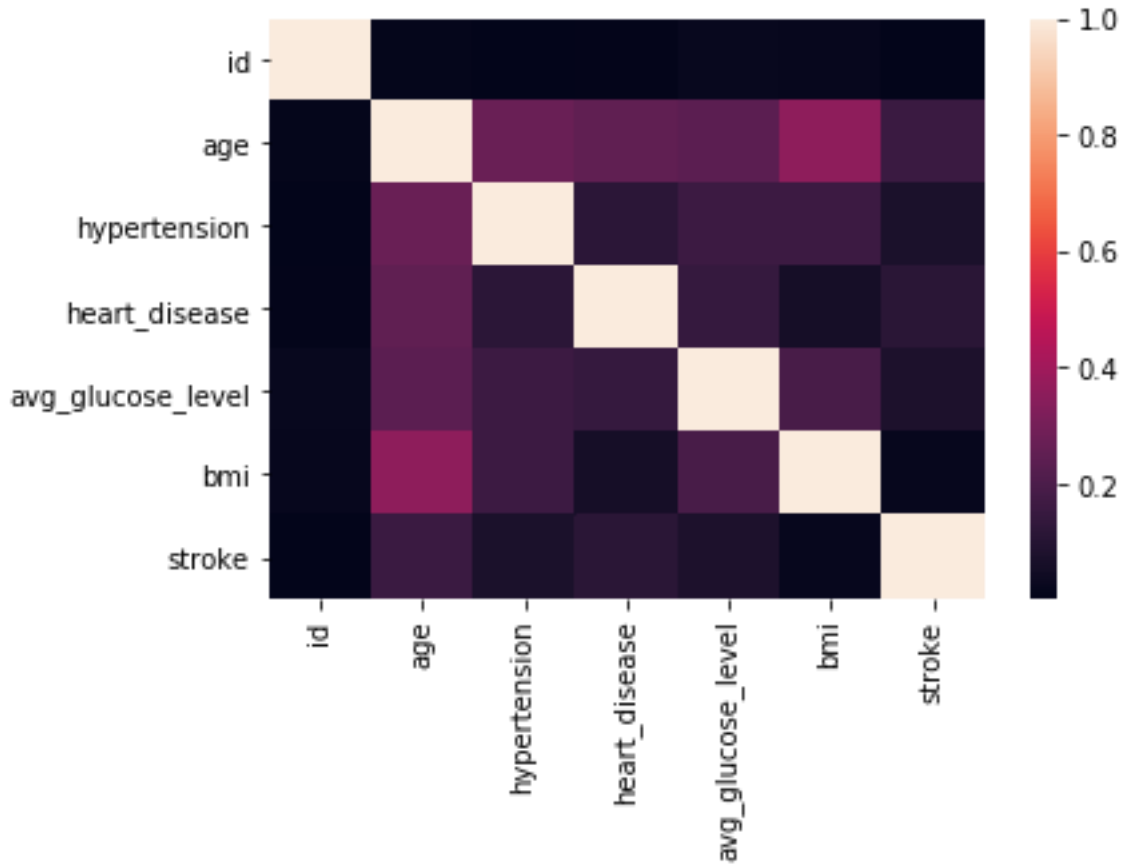


.



As the data is highly imbalanced, we couldn't find any noticeable relationship between the dependent and the independent variable .So we relay on our prior assumptions and decided to continue with all the independent variables We mainly used matplotlib and seaborn for making visualisations. The visualisations clearly uncovers the correlation between different variables in the data.

Heat map



From the correlation matrix we can find that age,glucose level,bmi and heart disease variable show positive correlation.

DATA CLEANING

Before applying the models we have to a prepare the data. Here we got some concerns. The “bmi” variable contains null values and are replaced with the mean (28.6) of “bmi” values. Another main issue is the “smoke status” variable where around 33 % of the values are null. It’s better dropping the column fully. As per our prior assumption smoke status and stroke are correlated. Hence we tried all models with two different datasets. One with smoke status value(replaced with mode ie never_smoked) and another without smoke status.

```
Classification report:
              precision    recall  f1-score   support

           0       1.00      0.98      0.99     10850
           1       0.00      0.00      0.00         0

    accuracy                           0.98     10850
   macro avg       0.50      0.49      0.50     10850
weighted avg       1.00      0.98      0.99     10850
```

Another problem with the data is the "imbalance" of the target variable.Above matrix shows the result of logistic regression model on the imbalanced data. Here we can see the precision value is 1.

The number of '1 ' is very low so that the preditions by the models may be biased. There are two methods to overcome the imbalaced data issue.ROSE(Random over sampling) and SMOTE(Synthetic minority over sampling). We here tried SMOTE method. The minority class got equal participation by filling with the synthetic data.

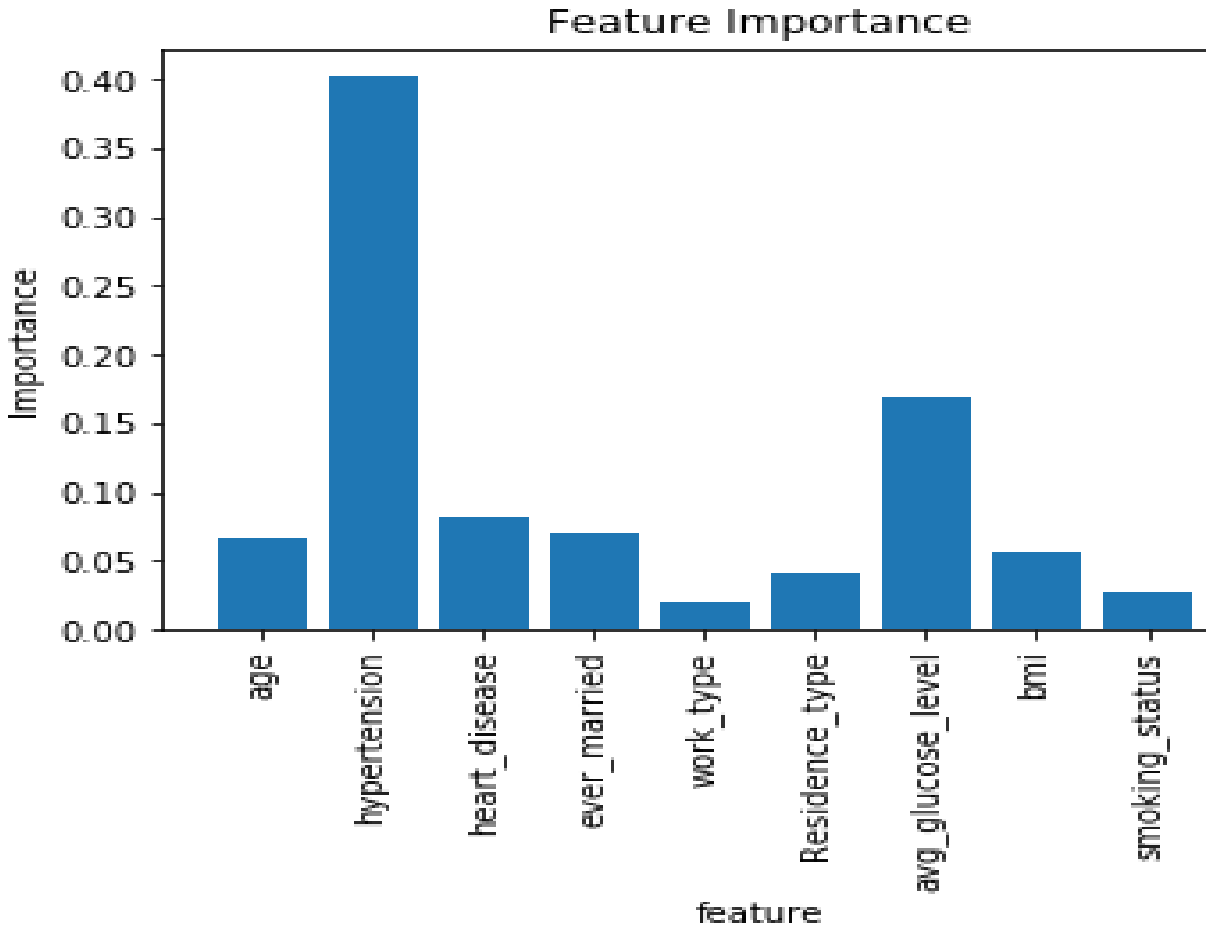
The data contains many categorical variables such as "gender", Ever married", work type ,residence type and smoking status. We performed label encoding for these variables.

RESULTS AND CONCLUSIONS:

We fit our data on different classification models: logistic regression, SVM, decision tree, random forest , KNN and all the results are compared .Almost all models performed well on the data. All the modes except logistic regression gave more than 90% precision on the predictions. For the heathcare field precision and accuracy of predictions are very important. The following table shows how different models performed on the data.

		Precision	Recall	Accuracy	F1_score	Confusion matrix
Logistic regression	With smoke status	0.7607	0.8124	0.7783	0.7857	[[9510 3268] [2399 10394]]
	Without smoking status	0.7608	0.8116	0.77814	0.7854	[[9514 3264] [2409 10384]]
SVM	With smoke status	0.9387	0.9903	0.9628	0.9638	[[11951 827] [124 12669]]
	Without smoking status	0.9320	0.9854	0.9567	0.9580	[[11859 919] [186 12607]]
Decision tree	With smoke status	0.9765	0.9817	0.9790	0.9791	[[12476 302] [233 12560]]
	Without smoking status	0.9760	0.9787	0.9773	0.9774	[[12471 307] [272 12521]]
KNN	With smoke status	0.8587	0.9933	0.9149	0.9211	[[10688 2090] [85 12708]]
	Without smoking status	0.8622	0.9899	0.9158	0.9216	[[10754 2024] [128 12665]]
Random forest	With smoke status	0.9977	0.9806	0.9892	0.9891	[[12750 28] [247 12546]]
	Without smoking status	0.9967	0.9789	0.9878	0.9877	[[12737 41] [269 12524]]

The above figure shows the feature importance of random forest classifier with the target variable.



We can see that the independent variables: hypertension, heart disease, average glucose level and age are having significance for the predictions made.

Marriage status, gender and work status variables can be ignored. surprisingly the smoking status variable didn't play a vital role as we assumed earlier.

Further investigation is needed in the case of smoking status variable, as we replced the null values with the MODE 'never smoked'.The lifestyle of a person is having great effect on the probability of occurring stroke. Those with hypertension is having higher chance of occurring stroke. We can further study the reasons for hypertension which may help to reduce the stroke occurrence.