STROKE PREDICTION

FUNDAMENTALS OF MACHINE LEARNING

PROJECT REPORT

Team Members:

1. ROSHNI SINGH - 19103034 (B10)
2. ABHISHEK KUMAR TAMOLI - 19103043 (B10)
3. RITIK RUSTAGI - 19103048 (B10)
4. UJJWAL SACHDEVA - 19104053 (B12)

Submitted to: Dr. Parul Agarwal

CSE & IT,

JAYPEE INSTITUTE OF INFORMATION TECHNOLOGY,
NOIDA

Contents

Problem Statement

Stroke is among major causes of death and long-term disability worldwide. As per reports, Stroke is the third leading cause of death and the principal cause of serious long-term disability in the United States. Accurate prediction of stroke is highly valuable for early intervention and treatment.

So, it is of great importance to predict the risk of having a stroke for better prevention and early treatment. This brief report presents my attempt to develop a machine learning (ML) model to accurately and quickly predict whether or not a person suffered stroke based on the Kaggle stroke dataset 1.

Motivation

As mentioned earlier, Strokes are a leading cause of death worldwide. So, accurate prediction of stroke is highly valuable for early intervention and treatment. So in this mini-project, we’ve considered some of the factors that might result in strokes like age, hypertension, bmi, smoking habits etc. Principles of machine learning are applied over large existing datasets. The prediction of outcomes in stroke patients may be useful in treatment decisions.

Literature available on the problem

1. An Introduction to Logistic Regression Analysis and Reporting.

Link:https://www.researchgate.net/publication/242579096_An_Introduction_to_Logistic_Regression_Analysis_and_Reporting

Abstract

The purpose of this article is to provide a set of guidelines for using logistic regression techniques. Tables, figures, and charts that should be included to comprehensively assess the results and assumptions to be verified are discussed.

Publication Details: September 2002 The Journal of Educational Research

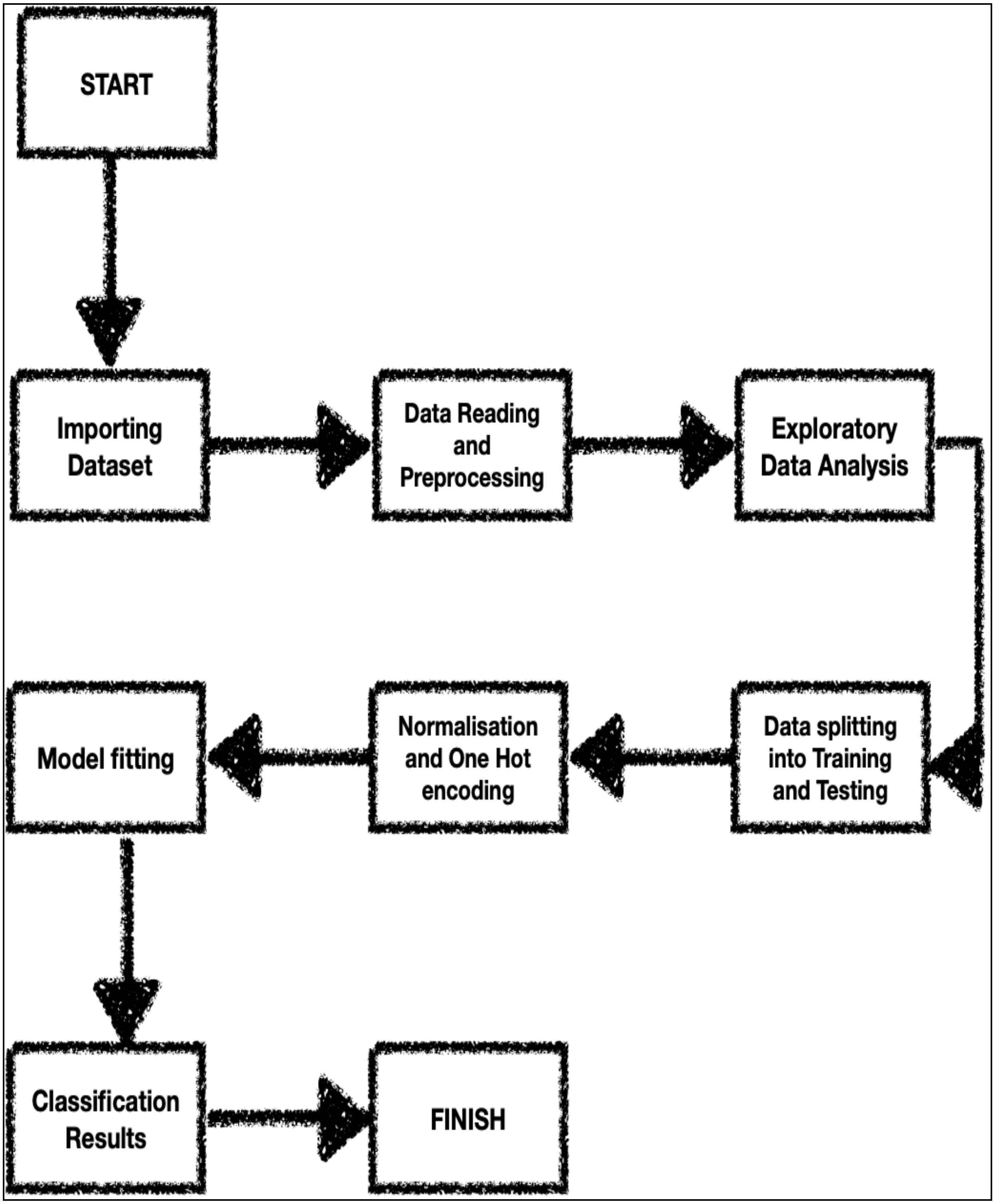
2. Stroke Risk Factors, Genetics, and Prevention

Link: https://www.ahajournals.org/doi/full/10.1161/CIRCRESAHA.116.308398

Abstract

Stroke is a heterogeneous syndrome, and determining risk factors and treatment depends on the specific pathogenesis of stroke. Risk factors for stroke can be categorized as modifiable and nonmodifiable. Age, sex, and race/ethnicity are nonmodifiable risk factors for both ischemic and hemorrhagic stroke, while hypertension, smoking, diet, and physical inactivity are among some of the more commonly reported modifiable risk factor.

Model used/proposed (flow diagram)


START
Importing Dataset
Data Reading and Preprocessing
Exploratory Data Analysis
Data splitting into Training and Testing
Normalisation and One Hot encoding
Model fitting
Classification Results
FINISH

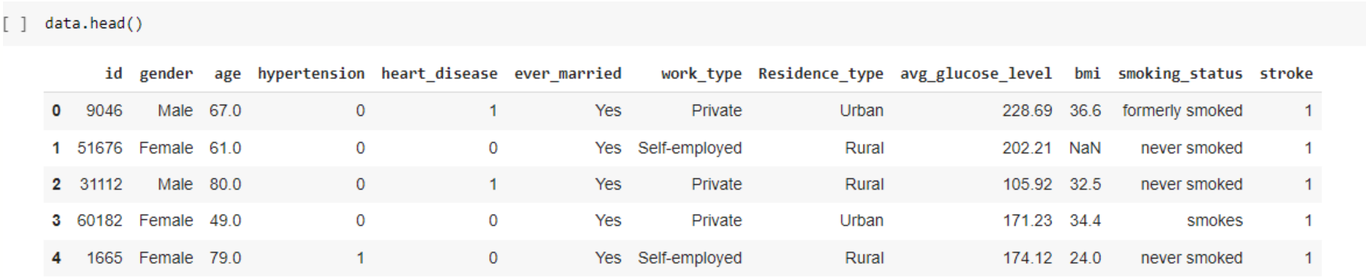
Dataset Description

The data contains 5110 observations with 12 attributes, ten of which are used as the input variables and described in Table 1. The remaining two variables are ‘id’ (excluded from our experiments) which corresponds to the patient's ID and ‘stroke’ which has a value of 0 (no stroke) or 1 (stroke) and is used as the output variable (target). The dataset exhibits a highly imbalanced class with 97.88% corresponding to class 0 and only 2.12% corresponding to class 1. There exist missing values of 3.37% and 30.63% in ‘bmi’ and ’smoking_status’ variables, respectively. In addition to providing performance benchmarks across various ML models, this report also examines which features are useful for stroke prediction.

ID	Corresponds to each unique record in dataset
GENDER	Classifies patient as Male or Female
AGE	Contains the age of each patient
HYPERTENSION	It’s a binary feature which reflect 0-No and 1-Yes
HEART_DIESEASE	It’s a binary feature which reflect 0-No and 1-Yes
EVER_MARRIED	Reflects the marital status of patient
WORK_TYPE	Classification of work type as pvt, govt or self-employed
RESIDENCE_TYPE	Reflects the locality of patient weather urban or rural
AVG_GLUCOSE_TYPE	Record of average glucose level in patient
BMI	Holds Body Mass Index
SMOKING_STATUS	Reflects weather the patient smokes or not
STROKE	Reflects the patient’s stroke count in the past

Implementation

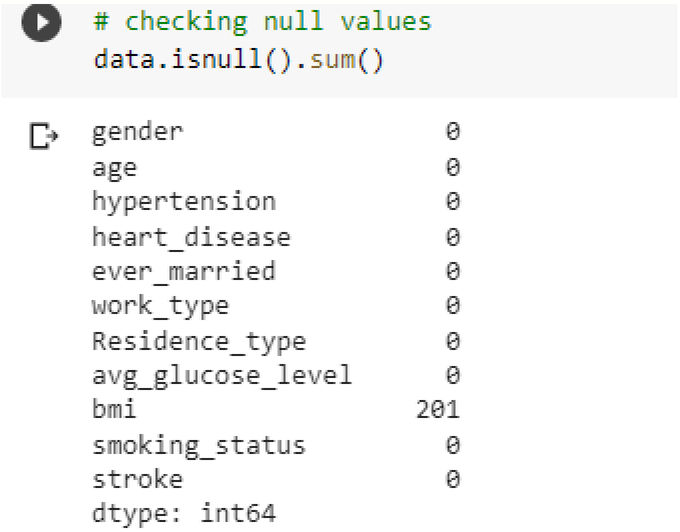
For the implementation, we have used the Stroke Prediction Dataset and imported it in the google colaboratory.



```
[ ]  data.head()
```

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1

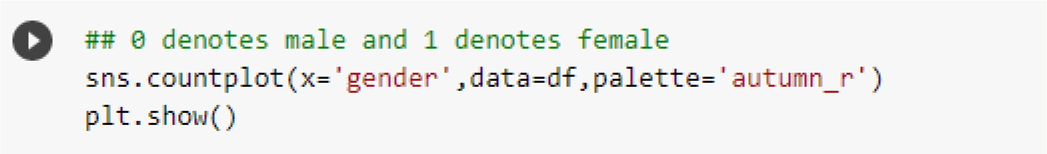
Id column was not necessary for prediction, so it was dropped from the dataframe. Then we checked for null values in the dataset and found that bmi column has 201 null values



```
# checking null values
data.isnull().sum()
```

```
gender                 0
age                    0
hypertension           0
heart_disease          0
ever_married           0
work_type              0
Residence_type         0
avg_glucose_level      0
bmi                  201
smoking_status         0
stroke                 0
dtype: int64
```

For filling the missing values, mean values of females and males have been replaced. After this, some exploratory data analysis has been done.



```
## 0 denotes male and 1 denotes female
sns.countplot(x='gender',data=df,palette='autumn_r')
plt.show()
```

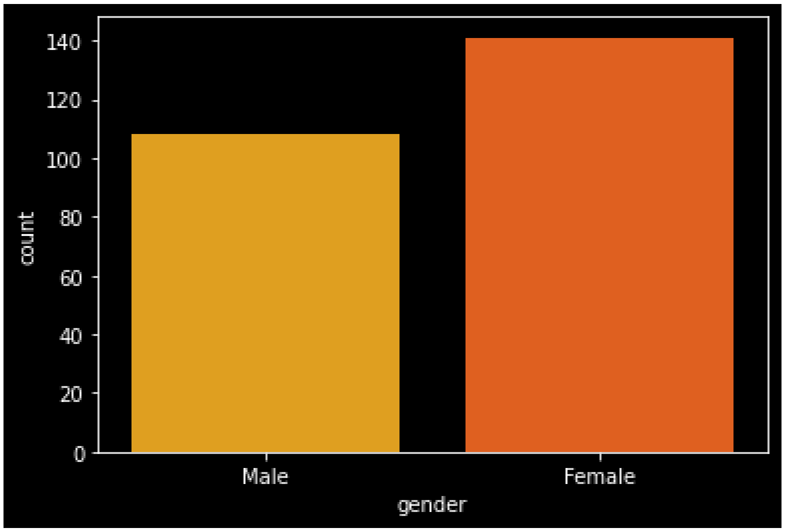


Fig 1

Fig 1 depicts the number of males and females who had strokes in the given dataset. This shows females have the highest cases.

Fig 2 shows the distribution of age groups who had strokes.

Fig 3 depicts that married people had the most number of stroke cases. And in particular females are higher in number in stroke cases.

Fig 4 shows that people who lived in urban areas had more cases of stroke than the people who lived in rural areas.

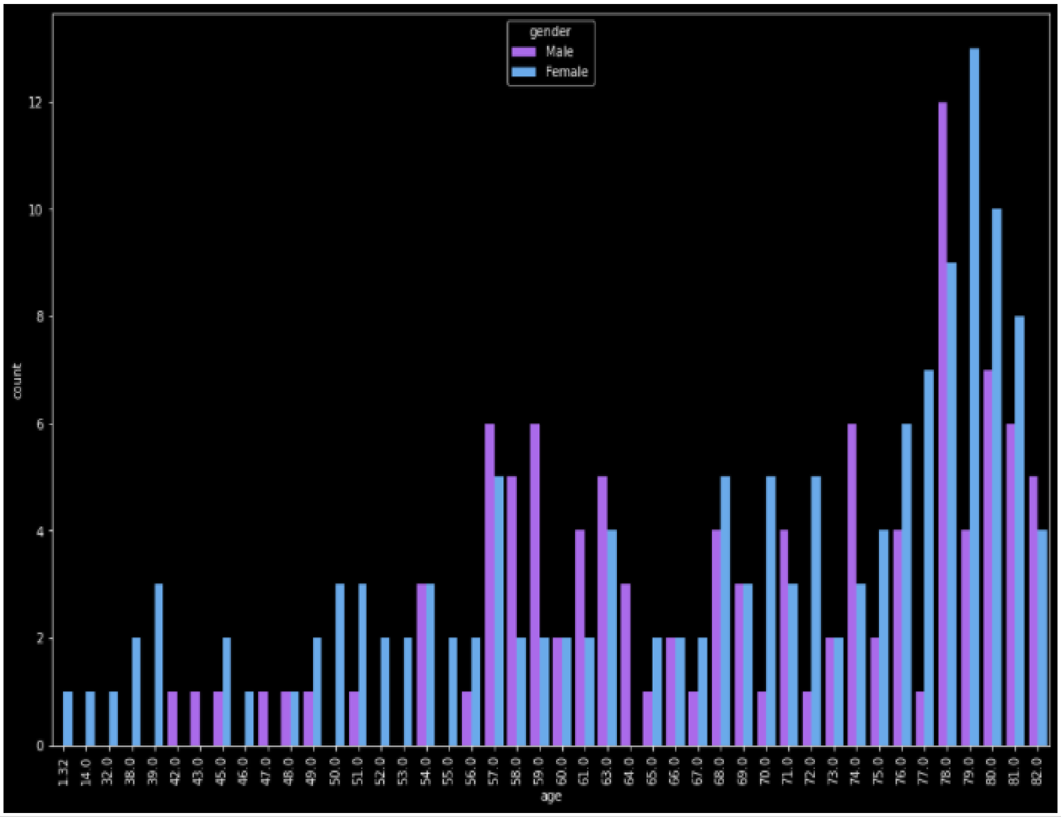

gender
Male
Female
count
0
2
4
6
8
10
12
age
1.32
14.0
32.0
38.0
39.0
42.0
43.0
45.0
46.0
47.0
48.0
49.0
50.0
51.0
52.0
53.0
54.0
55.0
56.0
57.0
58.0
59.0
60.0
61.0
63.0
64.0
65.0
66.0
67.0
68.0
69.0
70.0
71.0
72.0
73.0
74.0
75.0
76.0
77.0
78.0
79.0
80.0
81.0
82.0

Fig 2

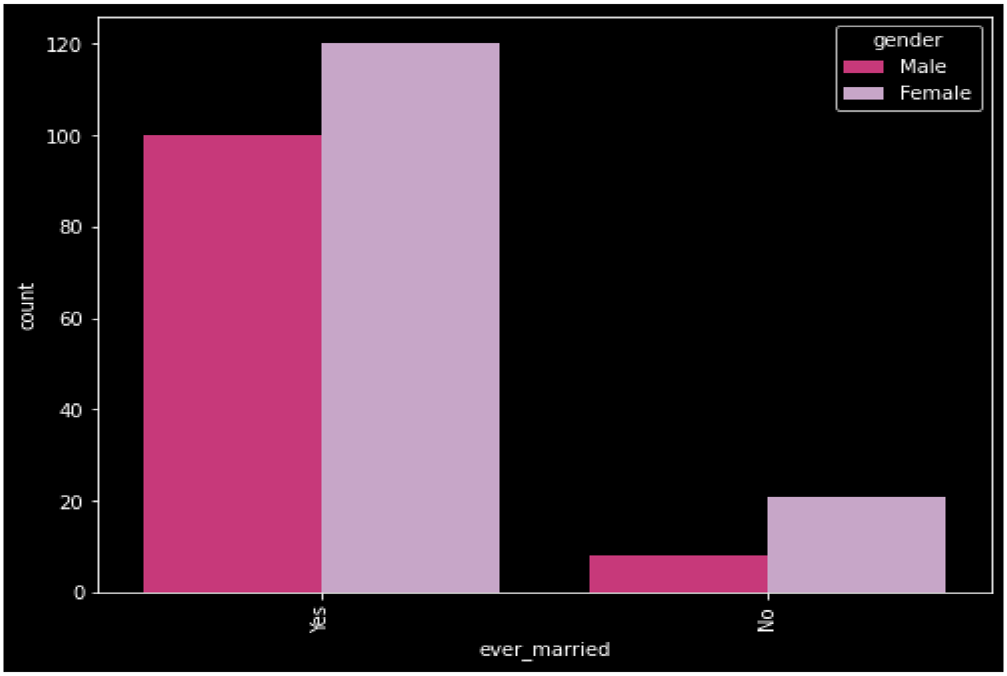

gender
Male
Female
count
0
20
40
60
80
100
120
Yes
No
ever_married

Fig 3

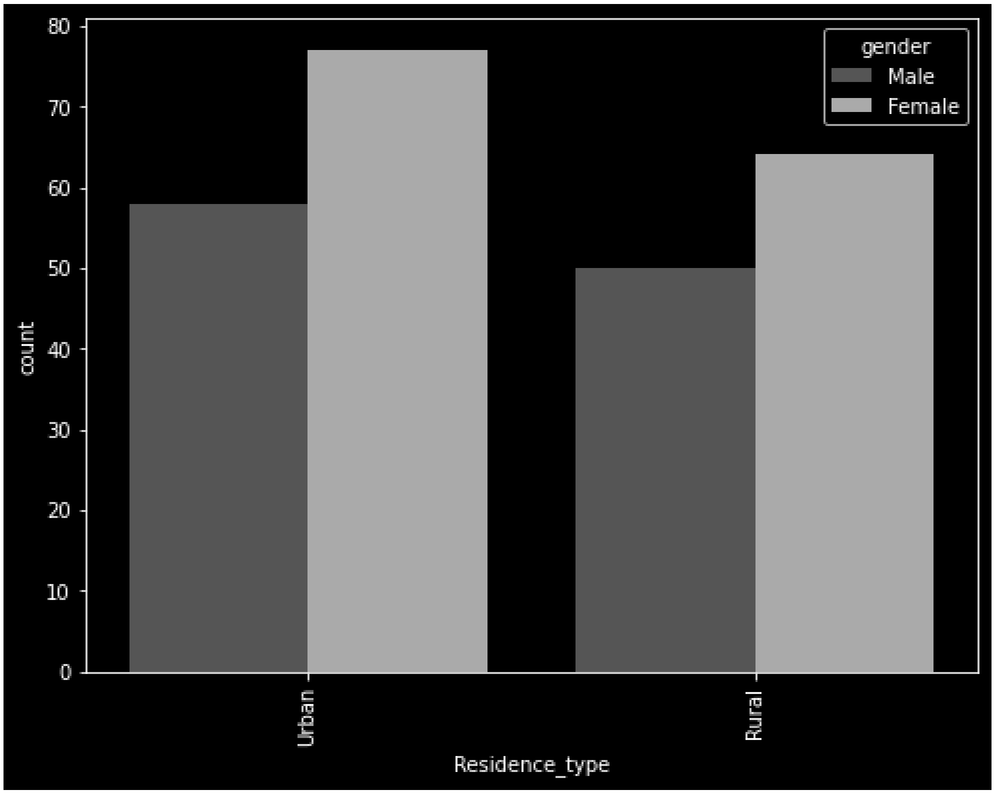
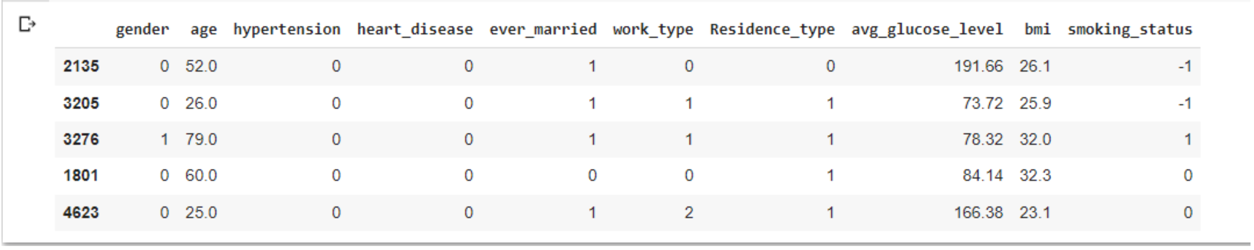


Fig 4

After doing exploratory data analysis, we defined the X (features) and y (target). Then the splitting of training and test data has been done followed by the preprocessing of data. Preprocessing of data here involves encoding all the categorical features into numerical values. Then normalization has been done on the dataset so that the data lies in a particular range.



	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status
2135	0	52.0	0	0	1	0	0	191.66	26.1	-1
3205	0	26.0	0	0	1	1	1	73.72	25.9	-1
3276	1	79.0	0	0	1	1	1	78.32	32.0	1
1801	0	60.0	0	0	0	0	1	84.14	32.3	0
4623	0	25.0	0	0	1	2	1	166.38	23.1	0

After the preprocessing and splitting of data, a logistic regression algorithm is applied and our model is fit to predict whether a person will have a stroke or not.

Results

Using the stroke prediction dataset, exploratory data analysis has been done. Afterwards Afterwards the dataset has been split into train and test sets and preprocessing is performed such as normalization and one hot encoding. After the preprocessing and splitting up the data into train and test sets, a logistic regression model is built. The accuracy attained is 95.49%.

Conclusion

This project demonstrates a method that uses logistic regression to predict whether a person can have a stroke or not by taking some input data from the user. The accuracy attained is 95.49%. Better accuracy can be obtained by using better classification algorithms such as support vector machines.

Contribution

Name	Enrollment No.	Batch	Contribution
Roshni Singh	19103034	B10	Model Building, Report
Abhishek Kumar Tamoli	19103043	B10	Exploratory Data Analysis, Report
Ujjwal Sachdeva	19104053	B12	Cleaning of dataset, Report
Ritik Rustagi	19103038	B10	Literature Analysis, Flow Diagram

References

1. https://www.researchgate.net/publication/242579096_An_Introduction_to_Logistic_Regression_Analysis_and_Reporting

2. https://www.ahajournals.org/doi/full/10.1161/CIRCRESAHA.116.308398

3. https://www.kaggle.com/fedesoriano/stroke-prediction-dataset