
DATA WARE HOUSING & MINING

Thejo Thanvitha Majety
AP19110010040
CSE-D

Problem Statement

- Stroke is a leading cause of death across the world and a major cause of severe disability in adults. The sudden death of living brain cells due to loss of blood flow or lack of oxygen flow to brain is almost fatal.
- So, here I am using the stroke prediction dataset here contains data like age, average glucose level in body, whether or not they had a heart disease or hypertension, etc of patients. The data is split into training and test data and the training data is used to create models that predict whether or not an entry from the test data will suffer a stroke.

Data set Introduction

	A	B	C	D	E	F	G	H	I	J	K	L
1	id	gender	age	hypertensic	heart_disea	ever_married	work_type	Residence	avg_glucose	bmi	smoking_st	stroke
2	9046	Male	67	0	1	Yes	Private	Urban	228.69	36.6	formerly sr	1
3	51676	Female	61	0	0	Yes	Self-empl	Rural	202.21	N/A	never smol	1
4	31112	Male	80	0	1	Yes	Private	Rural	105.92	32.5	never smol	1
5	60182	Female	49	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
6	1665	Female	79	1	0	Yes	Self-empl	Rural	174.12	24	never smol	1
7	56669	Male	81	0	0	Yes	Private	Urban	186.21	29	formerly sr	1
8	53882	Male	74	1	1	Yes	Private	Rural	70.09	27.4	never smol	1
9	10434	Female	69	0	0	No	Private	Urban	94.39	22.8	never smol	1
10	27419	Female	59	0	0	Yes	Private	Rural	76.15	N/A	Unknown	1

- The data set is about stroke prediction and it totally contains 5110 rows and 12 columns which are used to create a decision tree model.
 - Data source : <https://www.kaggle.com/>

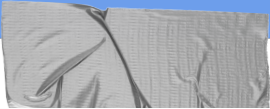


Steps Followed

- Data cleaning and Preprocessing
- Statistical Analyzing of the data
- Data selection
- Data transformation
- Model creating

Data Cleaning and Preprocessing

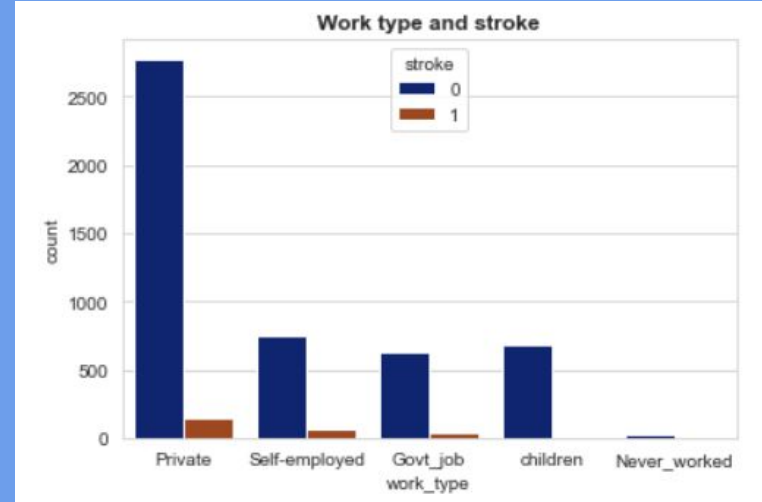
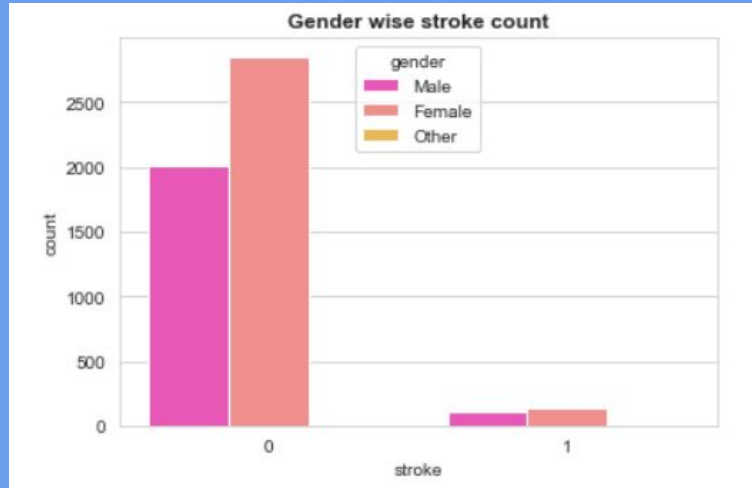
- We will be checking, if there are any null values and apply the different techniques such as Fillna, Dropna, manual filling and some statistical methods.
- Here I had obtained 201 missing values in BMI column and those are filled by
 1. Creating a function to classify bmi values into different groups.
 2. Based on the classification we will create box plots.
 3. Based on the averages obtained through box plots we will fill the average value.



```
df.isnull().sum()
```

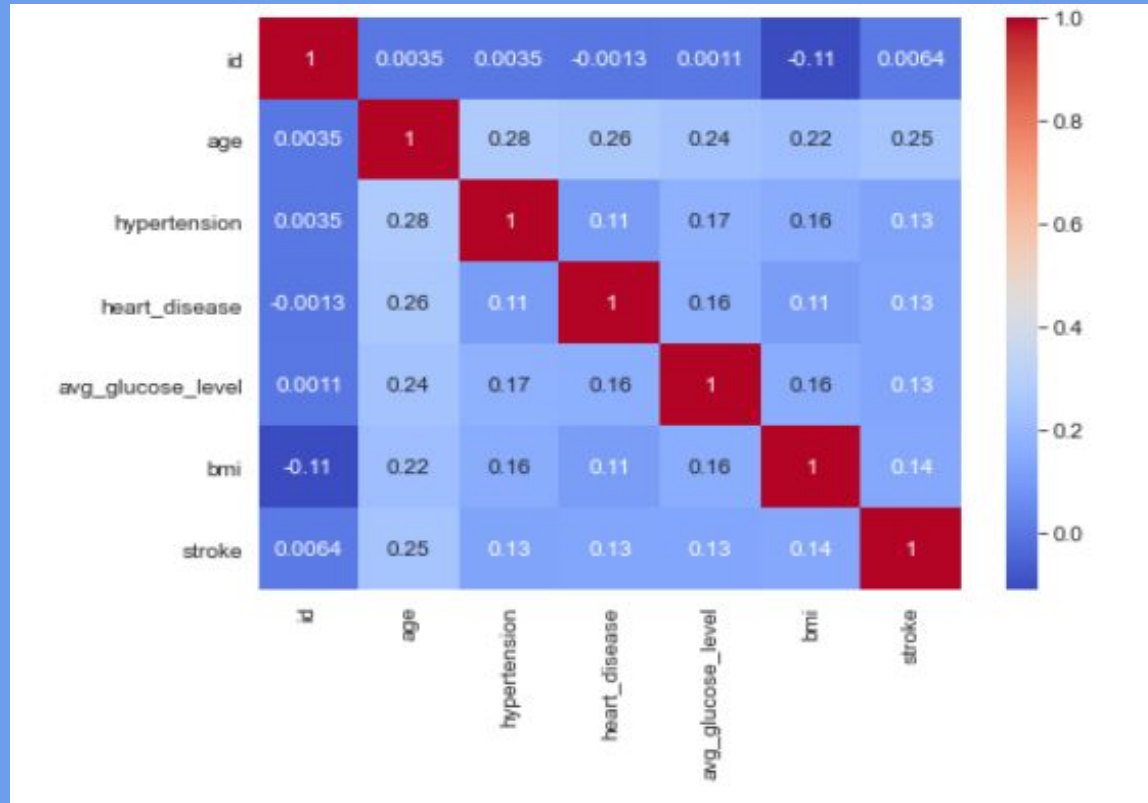
```
id                0
gender            0
age              0
hypertension      0
heart_disease     0
ever_married      0
work_type         0
Residence_type    0
avg_glucose_level 0
bmi              201
smoking_status    0
stroke            0
dtype: int64
```

Statistical Analysis



- As seen in the above pictures we will be performing few more statistical analysis to different columns to know the relation between them and to obtain the good accuracy while training our model

Heat Map to obtain Correlation



Data Selection

- Data Selection is the process where data relevant to the analysis task are retrieved.
- Here after completing the statistical analysis I had dropped few unwanted columns from the dataset.
- The data remained after data selection is

	age	hypertension	heart_disease	avg_glucose_level	bmi	stroke	Male	Other
0	67.0	0	1	228.69	36.6	1	1	0
1	61.0	0	0	202.21	105.0	1	0	0
2	80.0	0	1	105.92	32.5	1	1	0
3	49.0	0	0	171.23	34.4	1	0	0
4	79.0	1	0	174.12	24.0	1	0	0

Data Transformation

- Data transformation is the process of converting data from one format to another to make ease of testing and training of the model
- Here I had used dummies for categorical column -'gender' (that has strings)

Model Creating

- Based on the data obtained from the previous steps we need to decide which data mining algorithm is best suitable to test and train our model.
- Initially we need to import few required inbuilt libraries to start the test and training of our model from the scikit.learn
- We will be dividing the data into 2 parts(Train and Test) and it is a method to measure the accuracy the model. It is called Train and Test because we split the the data set into two sets : a training set and a testing set. 80% for training, and 20% for testing. You train the model using the training set.
- The algorithm I used for my model is Decision Trees

Decision Trees

- Decision Tree algorithm belongs to the family of supervised learning algorithms.
- The goal of using decision tree is to create a training model that can use to predict the value of target variable and here the data is continuously split according to a certain parameter.
- Here we are using function **Decision Tree Classifier** to build the tree.
- The accuracy of the model after implementing the algorithm is **90.7%**

