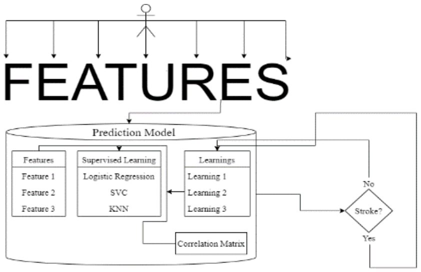
Machine Learning and Data Analysis approach for Stroke Prediction

Abstract: - This paper demonstrates about the serious disease which can be critical if not taken care of. This can be predicted based on our lifestyle, behavior, previous medical history, and mind stability. In this paper perform various data analysis method and build a perfect model using machine learning algorithm to predict about the significant illness.

Keywords: -***Machine learning, Data Analysis, Stroke Prediction, SVC, Logistic regression, KNN, Prediction.***

I. Introduction:-

In the UK there 100,000 cases per year out of which 38,000 causing death making it a leading cause of disability. According to Public Health England (PHE) one in six people will experience in their lifetime making a serious health concern. Major symptoms of stroke can sudden numbness in one or both arms or legs, confusion in speaking clearly or vision problem in one or both eyes and difficulty in walking in normal physical activities. NHS stated that the most common time (highest risk) is between 08:00am and noon. It is estimated that 7000 strokes can prevented, if taking precaution can care according to the doctors. This can save up to 2000 lives per year. When the nerves in the brain are trapped preventing the blood to pass through the arteries can cause stroke in that person. It can happen from many reasons where in the blood is inadequate to the brain. One of the examples can be hypertension and stress in life. Previous heart condition can also contribute for the causing of stroke. Stress in the life can happen because of many reasons such as money problem, marriage problem, workload, etc. Due to certain changes in our new generation's lifetime, it is more likely that stroke can happen to young and middle-aged people.



Hypothesis

We aim to find whether smoking habits and previous medical conditions according to our age will be the factors for experiencing a stroke.

Objectives: -

1- Identify the dependent variables and manipulate according to the requirement
2- Filtering out the irrelevant data so that the model is faster while predicting.
3- Use 2 or more techniques to find out which one has the better accuracy output.
4- Making the model with the perfect accuracy ready for real values apart from the present data.

II.Literature Review: -

The importance of the literature review is to build a profound knowledge of the crucial topic analyzed by several experts.

The first paper was 'Stroke risk prediction using machine learning: a prospective cohort study of 0.5 million Chinese adults' by Matthew Chun and member [1]. The aim of this research was to analyze the stroke prediction between population of China, so they collected around 500,000 surveys which roughly consisted 50% of men and women with no disability and was studied in terms of upcoming years. The paper was based on whether if the stroke happens between 0-3 years, 4-6 years, 7-9 years or after 9 years based on the collected data. The collected results where women in China tend to have higher number of chances for strokes then men but it also depends on the lifetime, the factor with higher number of people having strokes were also having other medical conditions such as diabetes, blood pressure and heart problem.

The second report is on ‘A Study of Features Affecting on Stroke Prediction Using Machine Learning’ by Panida Songram and Chatklaw Jareanpon. It is prediction study of people of Thailand, the data was collected from a hospital in which consisted of 500 values having various features such as 170 disease and smoking and age. The results in this paper were there were more number of stroke in people with habit of smoking, alcohol and no physical activity. People with abnormal cholesterol and blood pressure were also in the risk having stroke.

Third report which I went through is ‘Prediction of Stroke Using Machine Learning’ by Kunder Akash Mahesh and team. The paper presents some algorithm for predicting stroke such as Artificial Neural network, Bayesian Classifier and Decision Tree. The features which calculated were Glucose level, Gender, BMI, hypertension. The output of the model was in the favor of Navies Bayes classifier with better accuracy. Performance analysis shows that higher the AUC (Area under the Curve) better the model.

Final report shows, ‘Detection of Stroke Disease using Machine Learning Algorithms’ proposed by Tasfia Ismail Shoily and team. Here 4 algorithms Navies Bayes theorem, SVM, KNN, Random Forest Classifier. In this paper there 28 features used for prediction for good accuracy. Random Forest Classifier had the better accuracy amongst them and Navies bayes didn’t worked as presented in the paper. In the performance analysis in Random Forest classifier precision was up to mark with accuracy.

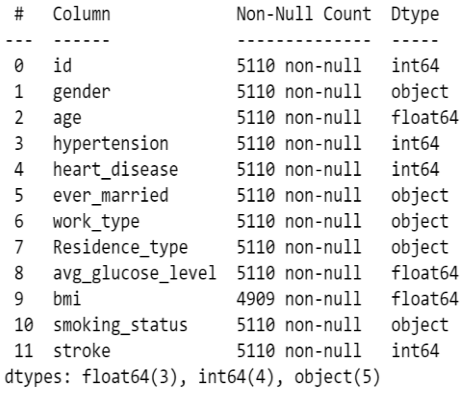
III. Data Processing and Management

The dataset we have used here consist of 12 features which are relevant for prediction of the model, but first we will need to clean the dataset and figure out about the null value.

1. Data Source and Description

The dataset was downloaded from www.kaggle.com .

The dataset is purely based on the health factor with 12 features for the prediction of the whether the person will have a stroke or not. If the person is going to experience stroke, then it will be 1 else 0. There are also some similar factors in the dataset such as heart disease and hypertension, they also have the similar values as in if hypertension or heart disease it will be 1 if *not* then 0. After importing our dataset in the notebook, we can see the features and their datatypes. Below figure shows the same.



```
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   id                 5110 non-null   int64  
 1   gender             5110 non-null   object 
 2   age                5110 non-null   float64
 3   hypertension       5110 non-null   int64  
 4   heart_disease      5110 non-null   int64  
 5   ever_married       5110 non-null   object 
 6   work_type          5110 non-null   object 
 7   Residence_type     5110 non-null   object 
 8   avg_glucose_level  5110 non-null   float64
 9   bmi                4909 non-null   float64
 10  smoking_status     5110 non-null   object 
 11  stroke             5110 non-null   int64  
dtypes: float64(3), int64(4), object(5)
```

Figure 1: Dataset Columns

As we can observe from the above figure we have 3 float, 4 int and rest object datatype in our dataset. Although after exploring each feature, we can see that ‘stroke’ feature is somewhat imbalance output. We can a greater number of ‘0’ then ‘1’.

```
<AxesSubplot:xlabel='stroke'>
```

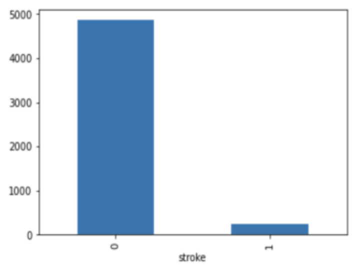
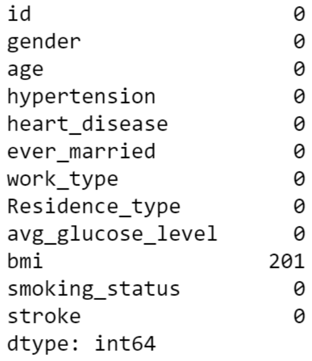


Figure 2: Plot bar for Stroke values

2. Missing value manipulation

While exploring the dataset we observed that there are some null values in the BMI values feature.



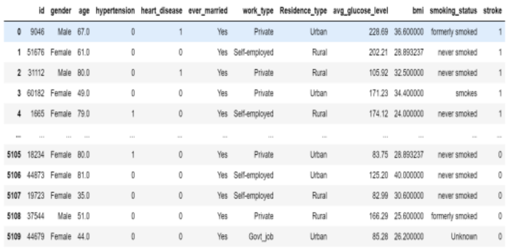
```
id                     0
gender                 0
age                    0
hypertension           0
heart_disease          0
ever_married           0
work_type              0
Residence_type         0
avg_glucose_level      0
bmi                  201
smoking_status         0
stroke               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

Figure 3: Null values in the dataset

Body Max Index is value which is unique for everyone.

BMI = $\frac{Weight(lb)}{[height(in)]^2} \times 703$

As we can see by the formula, we don’t have the values of weight and height of each person and this dataset has large number of values so with the given values we will just take the average value of the feature and fill it in place of null values.



	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	36.600000	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	28.893237	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.500000	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.400000	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.000000	never smoked	1
...	...	...	...	...	...	...	...	...	...	...	...	...
5105	18234	Female	80.0	1	0	Yes	Private	Urban	83.75	28.893237	never smoked	0
5106	44873	Female	81.0	0	0	Yes	Self-employed	Urban	125.20	40.000000	never smoked	0
5107	19723	Female	35.0	0	0	Yes	Self-employed	Rural	82.99	30.600000	never smoked	0
5108	37544	Male	51.0	0	0	Yes	Private	Rural	166.29	25.600000	formerly smoked	0
5109	44679	Female	44.0	0	0	Yes	Govt_job	Urban	85.28	26.200000	Unknown	0

Figure 4: Dataset after removing null values

Now all the null values are filled with the mean values from ‘bmi’ column. Also, to make the model run smoothly we convert the float values to int datatype. Also, we will be dropping ‘ID’ column as it is irrelevant to the prediction of the model.

3. Exploratory Data Analysis

In this part, we will explore every specific feature in the dataset and their relevancy to the model we are building at the end.

- Age

```
<seaborn.axisgrid.FacetGrid at 0x113be7d44c0>
```

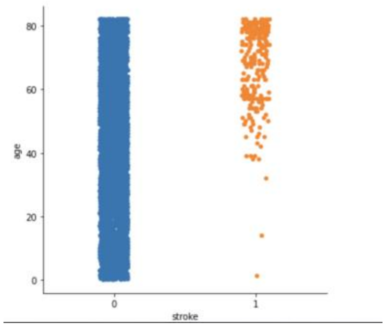


Figure 5: Cat plot for Age and Stroke

As this is the health dataset, age is the main factor considering the issues that come up as we grow old. As shown in our dataset we have different age category. We can see that almost each age category has at least some chances of having a stroke, but the old people are more likely to experience stroke with compared to young person. Therefore, it is an important factor in building model.

- Gender

The gender is significant factor for prediction as according to the health record it is shown then men are more likely to experience health conditions then women as medical surveys show men are more stressed and tensed.

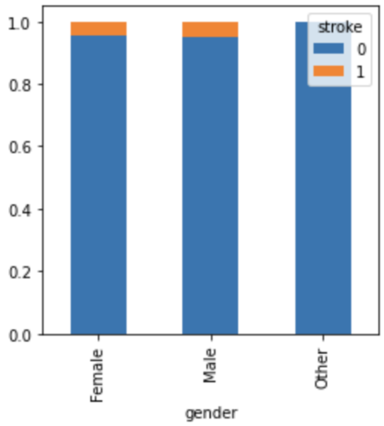


Figure 6: Gender and Stroke analysis

The figure above shows there is almost no difference between the two genders and other are not showing any signs of stroke.

- Hypertension, Heart Disease, Average Glucose level

This is very crucial factor for analyzing any medical condition, as per our dataset we only have the values in '0' and '1' for these two features. Hypertension is a disease in which person can through various problem such as high blood pressure which can someday lead brain disease. Similarly, any previous or current heart condition can cause the same. But glucose level is an independent feature and can also be an element for examining medical condition.

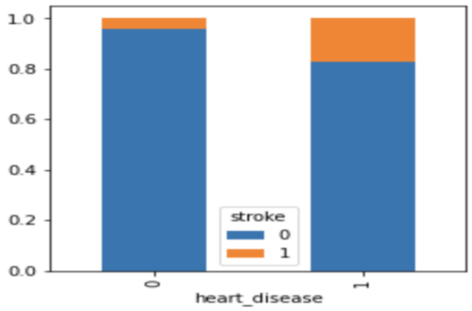


Figure 7:Heart disease vs Stroke

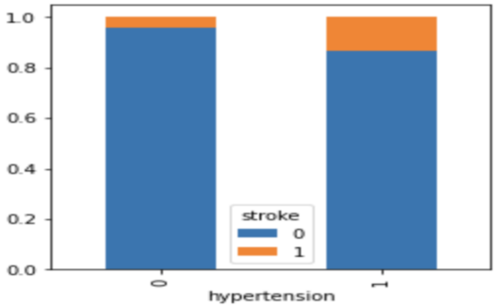


Figure 8: Hypertension vs Stroke

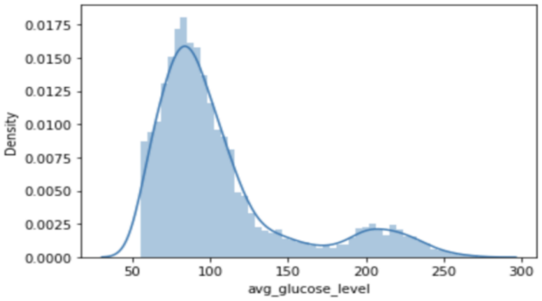


Figure 9: Independent Glucose level Graph

- Ever married and Work type

This feature is a part of bivariate analysis because they have less value in terms of input, therefore this type of graph will help us to get a good understanding of significance of the feature. as we compared in below Fig 10 Married tend to have a risk of stroke, and from Fig 11 self-employed, Govt job and private are more prone to stroke.

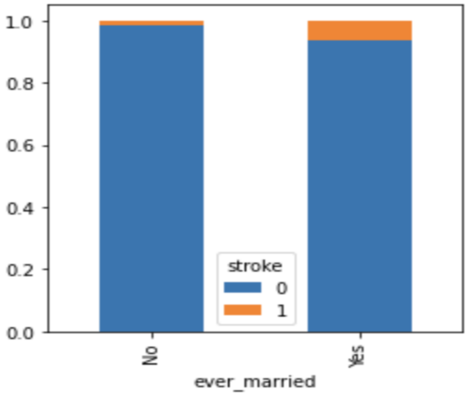


Figure 10: Ever married vs Stroke

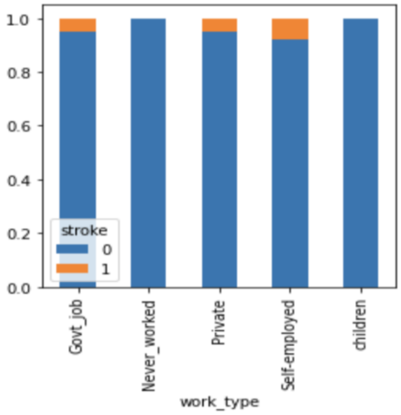
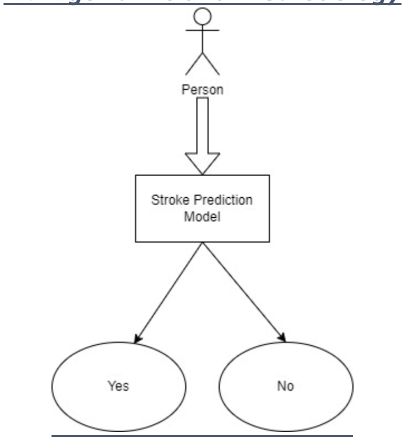


Figure 11: Working type vs Stroke

IV. Algorithms and Methodology



In this research paper we will use supervised learning algorithm in which we have known set of input and output, based on that model is created and it is trained and tested. Then the same model is used to get a reasonable prediction for the new inputs. We have used Logistic regression, Support Vector Classification and K – Neighbors Classifier. Here in all these algorithms, we got more 90% accuracy and cross validation scores. We have also used Two types of Data preprocessing methods, One hot encoding and Label Encoding and there was a little change in accuracy and confusion matrix between both of them.

In One hot encoding we get the columns values with less categories and divide them into two or more column depending upon them and converts them into 1's and 0's for that category.

On the contrary in Label Encoding for example if we have 3 categories in a column it will spilt and make only one column and assign them numbers according to the values of that category name.

Logistic Regression: - It is statistical model that is uses logistic function and gives output in binary variable whether yes/no or 0/1. It understands the relationship between dependent variable and one or more independent variables. In our dataset there are 6 dependent variable age, ever_married, bmi, gender, work_type and smoking status and there are 3 independent variables hypertension, heart_disease and avg_glucose_level.

Justification: - This algorithm classifies the values of experiencing stroke in reference to smoking status, previous heart disease, hypertension, and age. This technique just takes the input from the dataset apply the formula and gives outputs, it is easier to compute and gives good accuracy. We get two accuracies as we used two different methods on data preprocessing the output are shown in the program. As we can see there is slight difference between two methods.

KNN(K-Nearest Neighbor Classifier):- The algorithm is simple classifier which takes k number and searches that number of neighbor nearest to each other and so on it will choose every attribute and put them into the cluster. Therefore, at the end there will be only two cluster based on our dataset which will show us the output. Here in our model, we have selected 5 neighbors so that it will run faster and show better accuracy.

Justification: - Comparing it with other technique it takes way longer time to compute the algorithm and accuracy remains the same. It was a good choice as the classification problem helps us to understand dependency of the clusters upon the features. This technique is fast and more efficient with respect to output.

Support Vector Classification: - It is algorithm which uses hyperplane to separate the categories using high dimensional feature space to categorize the data even when the data is not linearly separable. Comparing with our dataset it will just categorize the stroke feature after calculating other features and will make a hyperplane(separator) between whether stroke will happen or not. It will the create a separator which is closest distance from the output variables. We have selected C as 1 because it is error term helps in optimization and we have used linear kernel which will just create a simple straight line.

Justification: - This technique classifies the hyperplane between 0's and 1's and computational time is more which makes it a little slow compared to logistic regression. But the accuracy is same as logistic.

***V.* Results:**

We have successfully used 3 algorithms on two data preprocessing methods and outputs were somewhat same. We have a 93.95% accuracy which is very precise in terms of prediction a foreign data. The confusion matrix was slightly different for each method.

Output of confusion matrix for Label encoding-

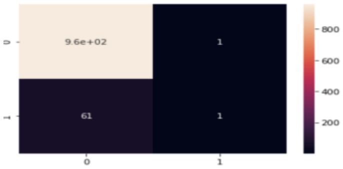


Figure 12:Logistic regression

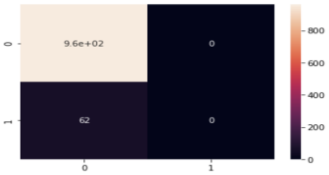


Figure 13:KNN classifier

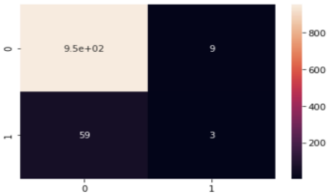


Figure 14: SVC classification

Output of Confusion using One hot encoding:

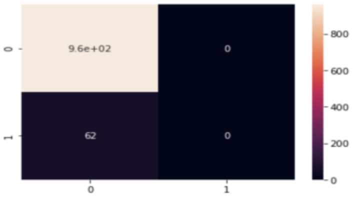


Figure 15:Logistic regression

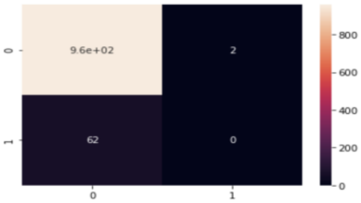


Figure 16:KNN classifier

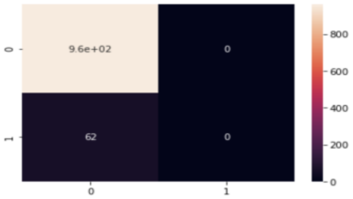


Figure 17:Support Vector Classification

***VI*.Challenges and Solutions**

The dataset consists of imbalance values of stroke feature if we had a greater number of feature the output would be more precise. There were more null values in the bmi feature as it is significant feature while predicting the accurate values, so to overcome this problem we just added the null values with mean values of BMI feature.

***VII*. Key findings:**

As we observe our dataset, we understand that there are features which needed more knowledge in terms of predicting a precise model such hypertension, there are various symptoms of this condition it should have mentioned in the dataset also heart disease like what kind of heart condition exactly or any major/ minor heart condition. This would have been a better dataset if it had person weight sometimes obesity can also be factor of causing stroke.

***IX*. Conclusion:**

The aim of the Hypothesis was satisfactory we were successfully able to predict that the previous health record and smoking habits can be a component of experiencing a stroke in our life and it can be more in number between older people than compared to younger people. We got good prediction accuracy for the large number of values. This model can also be efficient if we give any different values apart from the dataset.

Future Scope

This type of model can be used in real life situation where in predict person's chances of getting of stroke depending upon the person's day to day habits and previous medical records. This can help them take care so that there is no fatality in the life. Stroke is very critical disease if by any chance it's possible that we can create then we should never miss it. This model can be implemented in the real-life application if dataset is more precise.

***X*. Reference:**

[1] *'Stroke risk prediction using machine learning: a prospective cohort study of 0.5 million Chinese adults' by Matthew Chun and member*s.

[2] 'Prediction of Stroke Using Machine Learning' by KUNDER AKASH MAHESH, SHASHANK H N and team

[3]' Detection of Stroke Disease using Machine Learning Algorithms by Detection of Stroke Disease using Machine Learning Algorithms

[4]' Analyzing the Performance of Stroke Prediction using ML Classification Algorithms' by Gangavarapu Sailasya1 , Gorli L Aruna Kumari2 Department of Computer Science and Engineering GITAM Institute of Technology, GITAM (Deemed to be University) Visakhapatnam, Andhra Pradesh – 530045